Ville-Pekka Backlund

Temporal Percolation and Influential Nodes in Communication Networks

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 22.9.2014

Thesis supervisor:

Prof. Jari Saramäki

Thesis instructor:

Ph.D. Raj Kumar Pan

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Ville-Pekka Backlund

Title: Temporal Percolation and Influential Nodes in Communication Networks

Date: 22.9.2014 Language: English Number of pages: 7+58

School of Science

Department of Biomedical Engineering and Computational Science

Professorship: Computational Science Code: Becs-114

Supervisor: Prof. Jari Saramäki

Instructor: Ph.D. Raj Kumar Pan

A significant part of human communication is nowadays transmitted via electronic devices and applications which enable immediate contacts between individuals irrespective of location and time. An important side product of these media is the availability of large and detailed data sets on human communication that allow inferences to be made on the structure of the underlying social networks. The theory of temporal networks offers a suitable framework for studying time-resolved human communication both at the level of the whole system and at the level of individuals. This thesis studies three different real-world communication networks and addresses three questions.

First, percolation of temporal subgraphs constructed of consecutive communication events is studied. A phase transition from a fragmented to a connected phase and a percolation threshold is found in all networks. Emphasis is given to differences between static and temporal percolation, and on metrics that are of importance to the latter. Ensuring that the lifetime of temporal subgraphs spans the data interval was seen significant when determining the birth of a nucleus – a core group of nodes ultimately forming the giant temporal subgraph.

Second, the natural continuation of the first point is the study of those nodes who are responsible for the phase transition, *i.e.* nodes who are influential for the system. It is seen that the number of events and especially the degree of a node are good predictors of a node being part of the nucleus of the temporal network.

The third phase of the study focuses purely on the node level and addresses the influence of a node's actions within its local neighborhood. For this, a novel method of labeling the events of a temporal subgraph is presented. It is found that nodes with high degree and a large number of events are associated with larger temporal subgraphs.

Keywords: temporal networks, temporal percolation, influential node, complex networks, data analysis

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Ville-Pekka Backlund

Työn nimi: Aikariippuva perkolaatio ja olennaiset solmut kommunikaatioverkostoissa

Päivämäärä: 22.9.2014 Kieli: Englanti Sivumäärä: 7+58

Perustieteiden korkeakoulu

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Professuuri: Laskennallinen tiede Koodi: Becs-114

Valvoja: Prof. Jari Saramäki

Ohjaaja: Ph.D. Raj Kumar Pan

Merkittävä osa ihmisten välisestä kommunikaatiosta välittyy nykyään elektronisten viestinten välityksellä. Nämä viestimet mahdollistavat ajasta ja paikasta riippumattoman yhteydenpidon, sekä tuottavat suuria ja yksityiskohtaisia tietoaineistoja kommunikaatioverkostoista ja näiden kuvaamista sosiaalisista verkostoista. Aikariippuvien verkostojen teoria mahdollistaa näiden verkostojen tutkimisen sekä yksilöiden että koko verkoston tasolla. Tässä työssä tarkastellaan kolmea empiiristä kommunikaatioverkostoa ja tutkitaan erityisesti kolmea kysymystä.

Ensiksi, työssä tutkitaan peräkkäisistä puheluista tai viesteistä koostuvien aliverkostojen perkolaatiota. Kaikista kolmesta verkostosta tunnistetaan perkolaatiotransitio sirpaleisesta tilasta yhdistyneeseen tilaan sekä hetki, jolloin tämä tapahtuu. Työssä keskitytään vertailemaan kuinka staattisten ja aikariippuvien verkostojen perkolaatiotransitiot eroavat toisistaan, ja mihin erityisesti pitää kiinnittää huomiota jälkimmäisessä tapauksessa. Analyysin avulla voidaan todeta, että aliverkostojen elinikä on merkittävä käsite perkolaatiohetken määrittämisessä. Lisäksi pystyimme tunnistamaan koko verkostolle merkityksellisen aktiivisen ytimen synnyn.

Toiseksi, työssä tutkitaan ovatko tämän merkityksellisen ytimen solmut tunnistettavissa muista verkoston solmuista. Tulosten perusteella voidaan sanoa solmun kontaktien määrän ja erityisesti sen asteluvun selittävän hyvin solmun todennäköisyyden kuulua ytimeen.

Kolmanneksi, työssä tutkitaan solmujen käyttäytymistä ja merkitystä lähiympäristöilleen. Tätä varten kehitettiin menetelmä aliverkostojen kontaktien luokittelemiseksi. Havaintojen perusteella todetaan, että solmut joilla on suuri määrä kontakteja ja suuri asteluku esiintyvät suurempien aliverkostojen yhteydessä.

Avainsanat: aikariippuvat verkostot, aikariippuva perkolaatio, olennainen solmu, data-analyysi

Preface

This Thesis would not have been possible without the help and support of many.

First, I'd like to thank my supervisor, Professor Jari Saramäki, for the inspiration and contagious enthusiasm. Also, without the ultimate decision of employing me to the laboratory as a summer worker in 2010 this work would have never seen the light of day. I also want to express my gratitude to my instructor, Ph.D. Raj Kumar Pan, who has been my guide and collaborator in many projects and whose rigorous attitude towards science is second to none.

I'm honored to be given the change of working in the Complex Networks research group in the Department of Biomedical Engineering and Computational Science of Aalto University. The multicultural atmosphere and coffee breaks, created by the great people, are something that I certainly will miss. The mastermind behind this all, Professor Kimmo Kaski, deserves acclaim.

To my closest colleagues in the legendary room F340 – Rainer, Pauli and Joonas – I'd like to thank you for sharing the highs and lows of science and studies, and express my apologies for interrupting your work by commenting on recent news.

Especially, I want to thank my parents Jukka and Marketta who helped me in my studies in the best way imaginable: by staying completely out of them since the first day of elementary school.

Most of all, I'm grateful to my dear Marianne who helps me to be a better person.

Espoo, 22.9.2014

Ville-Pekka Backlund

Contents

Symbols and Abbreviations

Symbols

$\mathbf{A}$	adjacency matrix
B_i	burstiness of node i
C	global clustering coefficient
C_i	local clustering coefficient of node i
$d_{i,j}$	shortest path distance from node i to node j
$\mathcal{E}$	set of events of a temporal network
E	number of events in a temporal network
E_i	number of events of node i
e	a single event of a temporal network, $e \in \mathcal{E}$
G	a network
G_T	a temporal network
k_i	number of neighbors of node i, *i.e.* degree of i
L	set of links in a static network
l	mean shortest path distance between all pairs of nodes of a network
m	number of links in a network
N	number of nodes in a network
T	time difference between the first and the last event in a data, *i.e.* data duration
Δt	parameter for TSG search, limits the maximal time allowed between adjacent events
Δt_c	temporal percolation threshold
V	set of nodes in a static or temporal network
w	weight of a link

Operators

$\bar{O}$	average of quantity O over the nodes of the network
$\lvert O \rvert$	size of set O in unique nodes

Abbreviations

ER	Erdős-Rényi network
LCC	largest connected component
RTS	random time shuffle
SMS	mobile phone short message
TSG	temporal subgraph
TSGEL	temporal subgraphs with event labeling

1 Introduction

Network science has helped us to extract relevant information from a great variety of complex systems. The reason behind its success is that it provides a simple approach for studying systems and their differences: just define the elements, *nodes*, of the system and represent their mutual interactions with *links*. For instance, network science has been successfully used in studying metabolical networks [1], information networks such as the World Wide Web [2], and especially, social networks [3–6].

Sociologists were among the pioneers of using – and developing – network science already in the 1960s. However, social networks are now more relevant than ever. This is due to the immense advancements in communication devices and the explosive growth in their use we have seen during the last few decades. Eventually these lead to (only exaggerating slightly) *status quo*: any two people wherever can connect with each other whenever.

In particular, people's inherent need to communicate and their desire to be in contact with each other has lead to mobile phones, email, and other means of electronic communication, as well as services built specifically for social networks, such as Facebook. For network science these media are significant because of their ability to keep records of everything that takes place on them. Especially, such electronic data sets enable us to perform studies and test hypotheses at the scale of large groups or even populations, not just at the level of individuals or small groups where sociologists were long limited to.

Treatment of these immense data sets requires computational power, but in order to utilize all the details of the data, a more elaborate representation of the underlying social network itself is needed. In particular, such sets of data are in fact composed of streams of interactions events (calls, emails) that contain information on social dynamics on multiple time scales. Hence the traditional view of social networks as static entities representing a "snapshot" of the state of affairs at some particular point in time is no longer sufficient. This leads to the framework of *temporal networks* [7]. With the dimension of time, we are able to account for both the structure *of* the underlying networks and the interaction *events* happening *on* the network.

Studying the occurrence patterns of events of empirical human communication networks reveals information both at the system and individual level. Particularly, at the network level, it is interesting to see if there exists a characteristic time scale within the system which limits all phenomena, *i.e.* to study the *temporal percolation* of the network. At the other extreme of the size scale lay the nodes representing individuals – the definitive building blocks of any social network. Clarifying the level of influence of an individual on the system and on their local neighborhood – given a restricted amount of data – has applications beyond basic research.

1.1 Objectives and Scope

The concept of percolation is related to whether a system is connected at global scale or fragmented to multiple smaller parts. Essentially, this reveals if the underlying framework offers a substrate for global phenomena. In the social context, one example of such is information spreading, and percolation helps to predict whether a piece of information can reach the majority of individuals in a social network.

As motivated above and stated in the title, **studying temporal percolation in a communication network** is the main objective of this Thesis. As percolation in temporal networks has been little studied, this requires defining novel concepts and metrics. These are put into use when inspecting three empirical communication networks constructed from three different sources, namely mobile phone calls, mobile phone short messages (SMS) and emails. We are especially interested in whether one can detect the percolation threshold – a specific point at which the system becomes connected – and thus the birth of the *nucleus* of the giant temporal subgraph percolating the system, that is, an influential group of nodes who are responsible for this transition.

The first immediate follow-up question is whether the nodes forming this influential nucleus are different from the other nodes? We study this by choosing a node property and observing how it affects a node's likelihood of participation in the nucleus. The explanatory properties are chosen specifically so that they only utilize local information on nodes, "local" meaning the network structure immediately surrounding the node. This restriction is based on the nature of social networks: most processes of social influence and information transmission are rather local and restricted to small network neighborhoods. Also, the applicability of the results is enhanced if no computationally expensive network-level properties for nodes need to be calculated.

Second – in a more literal meaning of the word "influential" – we are interested how the actions of a single individual node affect its local neighborhood. To study this, we construct a novel approach of labeling the events of the nodes. The method enables us to study the role of a node with respect to its neighbors and their events.

These two questions are combined to the second objective of this Thesis. That is, we want to investigate **which nodes in temporal networks are the most influential regarding temporal connectivity and flow of information.**

The pursuit for the objectives begins in Chapter 2 with an introduction to network science and especially to the temporal networks framework. We describe the basic definitions and tools, and also highlight some important findings. Emphasis is given to the concept of percolation. Next Section is devoted to discussing the characteristics of social networks and human communication networks. The last Section of Chapter 2 is dedicated to representing related work on influential node identification on both static and temporal networks and clarifying our targets on the matter.

In the following Chapter 3 we introduce the methods and the data sets used in this

Thesis. Next, in Chapter 4, we will report the results of the studies. As we have two main objectives, they both have their individual discussion Sections after the corresponding results are represented. Finally, this Thesis is finished with a general summary and conclusions.

2 Background

In essence, network science is a multidisciplinary field combining ideas from mathematics, physics, computer science and sociology. It has gained attention because of a straightforward principle: it is simple, and especially because it works. In this Chapter we go through the principles of network science, temporal networks, percolation, special characteristics of social and communication networks, and related work on identifying the influential nodes of a network. An informed reader can skip the familiar parts without losing consistency, whereas additional information can be found in the comprehensive works of References [8, 9].

2.1 Introduction to Network Science

Although some of the key ideas of network science are already centuries old, this umbrella term for the whole field has become established in the 21st century.[1] A seminal article by Watts and Strogatz on small-world networks from 1998 can be seen as a breaking point for the whole field [10]. The biggest contribution of the article was that the differences and similarities of real networks constructed from completely different sources (social, biological and technological) could be demonstrated with simple tools that have now become parts of the canonical toolbox of network scientists.

The first step of using the network framework is constructing the network (also called a *graph*) from data representing real-world phenomena. In principle this is very easy, we just need to define who are the elements of the system and represent them as *nodes* (or alternatively called *vertices*) of the network. Then, the interactions between the elements of the system are represented in the network by connecting the nodes with *links* (*edges*). Note that there is no unique way of defining the nodes and links; often, we can construct multiple different network representations from the same data. Formally, this step consists of constructing the set of nodes V and the set of links L which together define the network $G = (V, L)$. If not defined otherwise, the network is assumed to be unchangeable. This can be emphasized by calling it a *static* network.

The next step is choosing a suitable way to represent the network in practice. The most traditional one is to use the *adjacency matrix*, which is defined element-wise as

$$A_{i,j} = \begin{cases} 1 & \text{if there is a link connecting node } i \text{ to node } j \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Also, depending on the system that the network represents, we need to define whether it is *directed* or *undirected* (*bidirectional*). In a directed network, the interactions represented by the links are considered to be directional, whereas in the

[1] For a classical example, see the problem of the seven bridges of Königsberg `http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg` (11.6.2014.).

undirected case $A_{i,j} = A_{j,i}$ holds for all i and j. It is also quite common that the interactions depicted as links are not equally strong. Their strength differences can be included in the representation by assigning an arbitrary weight $w_{i,j}$ to each link.

Since real-world networks are usually sparse, the adjacency matrix representation is inefficient for computational purposes as the large number of zero-valued elements requires a lot of memory. Often it is replaced for instance by a neighbor list, where the format `node_i:node_j,node_m` describes first the node in question and then the nodes it is connected to. Figure 2.1 clarifies these concepts.

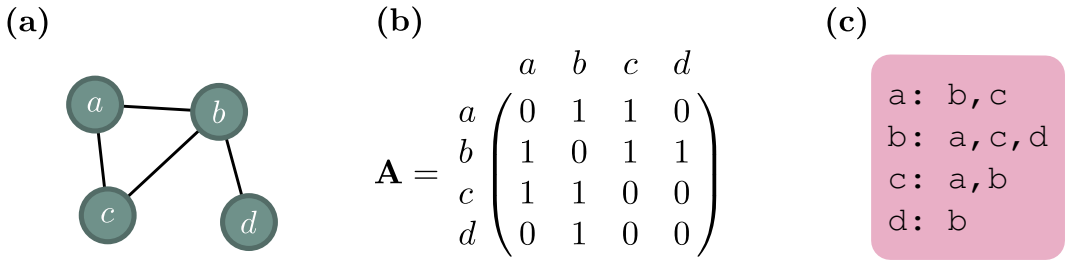


Figure 2.1: An example of an undirected network with four nodes and four links (a) and the corresponding adjacency matrix (b) and neighbor list (c) representations.

Now that we know how to represent networks, we can move on to see what can be done with them. For this, we need metrics which describe the networks.

On a general level, metrics are applied to measure something. They have descriptive power just by themselves, but often we want to state something about the relative strength of some phenomenon and thus we need a reference point to compare to. An important baseline for all the network measures is the Erdős-Rényi random network model (ER). The most common version of it is constructed by taking N nodes, going through all the possible links between the node pairs, and using probability p to define whether a link exists or not. This is denoted as the $G(N, p)$-network ensemble. As a more elaborate reference, the so-called *configuration model* is often applied. In this model, nodes can have arbitrary numbers of links – *e.g.* exactly the same as the nodes of some real-world network – but the networks are otherwise maximally random (see *e.g.* [8]). Then, if a real-world network has a property that differs from the same property of the reference network, we know that the difference originates in the way the real network evolved or was constructed. The concept of using or creating the proper reference is omnipresent in every aspect of network science.

The simplest network metrics include for instance the number of nodes in the network, denoted conventionally with N, and number of links m. However, these contain no information on *how* exactly the nodes are connected. In the next three Sections we introduce more elaborate concepts for studying networks and the phenomena they reveal, moving gradually from small to larger scales.

2.1.1 Degree and Scale-Freeness

The simplest way to study how the links are distributed between the nodes is to calculate the number of links each node has, denoted with k_i and called the degree of node i. Of course, in the case of directed networks, we can consider the incoming and outgoing links with in-degree and out-degree, respectively. The established practice is to call nodes connected by a link *neighbors* of each other, and thus the degree of a node equals the number of its neighbors.

Exploring networks at the level of degrees already gives some interesting results. For instance, we notice that in social networks high-degree nodes are usually connected to other high-degree nodes, in other words, popular people know popular people. This phenomenon is called *assortativity* [11].

Even more information of the network is revealed when studying the degree distributions, *i.e.* the probability density functions of degrees. It has been shown that often the distributions have broad tails, meaning that while most nodes have small degrees, there are always some nodes with very large degrees. Such networks where the degree distribution p_k, that is, the probability that a node has a degree k, follows a power-law ($p_k \propto k^{-\alpha}$) are called *scale-free* [12, 13].

Nodes with many connections, aptly called *hubs*, have an important role in the function of the networks. For instance, if the hubs are removed from the system, the network breaks down easily: think of the consequences of closing down the Heathrow and JFK airports at once.

2.1.2 Clustering, Paths and the Small-World Phenomenon

Moving from individual nodes to larger neighborhoods, the next question is how links relate to one another.

One important and simple metric describing this is the *clustering coefficient*. It has both global and local versions. The global one is defined as

$$C = \frac{3 \times (\text{number of triangles in the network})}{\text{number of connected triples of nodes}}, \tag{2.2}$$

where a *triangle* means a set of three nodes connected with three links, and a *triple* a node with two distinct neighbors. The global clustering coefficient is also called *transitivity*, which reveals the purpose of the metric better: if we consider that the link connecting two nodes is a relation, then transitivity for that relation means that if nodes i and j are connected and nodes j and k are connected, then also nodes j and k are connected by a link (*i.e.* $A = B, B = C \Rightarrow A = C$). In social context this means that you, and a friend of your friend, are connected. Thus, if a network has a transitive tendency between the links, then the clustering coefficient displays higher values. It has been shown that this is true for most real-world networks, especially for social networks [9].

The local clustering coefficient is a similar metric as the global version, with the difference that it considers only the local vicinity of each node individually, defined as

$$C_i = \frac{\text{number of triangles connected to node } i}{\text{number of pairs of neighbors of node } i}. \tag{2.3}$$

C_i is undefined if the degree of a node is less than two. The usual convention is to handle these cases by defining $C_{i,k_i<2} \equiv 0$ or discarding them from the analysis. With this definition clustering can be set against node-specific metrics, such as degree, in order to find out possible dependencies or their lack. To conclude, both clustering coefficients are measures of the density of triangles in a network. Usually the clustering coefficients are calculated for directed networks by considering the links undirected, though taking the directionality into account is possible if specifically needed.

Triangles are constructed with three links connected in a certain way, but we can also study an arbitrary chain of links. This brings us to the very important concept of a *path*: a sequence of nodes connected by links. To make the definition more usable and intuitive, we usually restrict that the same node or link can not exist in a path twice (called a *self-avoiding path*). Note that as paths are sequences, the links in them have an order. In addition, in directed networks the order must follow the direction of the individual links (*i.e.* a one-way street). The length of a path is the number of its links. In short, the path length gives us a distance metric for networks.

There are networks where some pairs nodes do not have a path between them, *i.e.* their distance is infinite. This is possible for undirected networks only if the network consists of multiple *components* that do not have any links between them (a component is defined as a subset of nodes where all nodes can be connected via some path). Also, in directed networks, a path may only exist between nodes i and j but not the other way around. Clearly, the possible structural separation in the form of components is a major bounding factor for everything in networks, and thus the size of the largest connected component (LCC) of a network, measured in nodes, is often studied.

When paths are known, one can study how compact a network is by asking how far away a randomly chosen node is from another randomly chosen node? To answer this, we need to study the length of the *shortest path* between two nodes i and j, also called a *geodesic path* and denoted with $d_{i,j}$. The mean shortest distance between all possible pairs of nodes of a network, l, is defined as

$$l = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \neq j} d_{i,j}, \tag{2.4}$$

where the links are now assumed to be undirected and the network to consist of one component only. The mean distance greatly affects the dynamics taking place on networks. Consider for instance a data packet traveling in the Internet: if the mean

shortest distance is small, the packet goes through a smaller number of routers and thus reaches its destination faster.

In real-world networks, the mean shortest distance has been seen to be surprisingly short. This phenomenon has a representative name: networks with short typical distances are called *small-world* networks [10].[2] As already motivated with the Internet example above, the small-world property allows information, or any other diffusion process such as epidemics, to spread effectively. In the social context the effect means that all the humans on Earth are connected through a short chain of acquaintances. This was the remarkable result obtained by Milgram in 1967 [4] and more recently seen in a planetary-scale Internet-based communication network [14].

2.1.3 Mesoscopic Level

The next logical step in the analysis of networks is to study groups of multiple nodes to reveal mesoscopic structural properties of networks. At its simplest, this amounts to studying a subgraphs of the original network $G' = (V', L')$ by taking subsets of nodes $V' \subseteq V$ and links $L' \subseteq L$. However, without proper rules of choosing the subgraphs, not much can be gained. One such rule is to take all disconnected subgraphs, leading again to the definition of a component. As seen next, there are other rules that split networks into parts that may overlap.

As seen above, clustering coefficients measure the density of triangles in a network. However, networks may have recurrent patterns of groups of nodes of arbitrary shape or size. This leads to the concept of a *motif*: a pattern of nodes and links which is overrepresented compared to a randomized reference network [15]. For example, it was seen that in food networks, where a directed link exists from species X to species Y if X feeds on Y, a chain of three nodes is much more common than would be expected at random.

A widely seen property of networks is that they have groups of nodes that are densely linked between each other but have only few links connecting them to other such groups. This phenomenon leads us to the concept of *community* [16]. For instance, scientists tend to collaborate with other scientists in their field, and this separation of disciplines is seen in collaboration networks where links connect people who are authors in a joint paper. There are only few scientists who participate in interdisciplinary collaboration, thus connecting the different fields (communities) [17]. Contrary to motifs which measure the occurrence of predefined patterns in the network, communities may come in a multitude of sizes and shapes, typically determined by some algorithm that optimizes their boundaries. Partly because no single perfect definition for a community exists, and because of the fact that in most cases it is hard or even impossible to know what the real "ground truth" communities which should be found are, there is a plethora of different community

[2] A more rigorous definition for the small-world effect requires that the mean shortest path distance scales logarithmically as a function of the size of the network, $l \propto \log N$.

detection algorithms. Some well-performing and popular algorithms are Infomap [18] and the Louvain method [19] but still, community detection must be utilized with prudence [20].

2.1.4 Dynamical Processes and Static Percolation

Because many real-world processes take place on networks, networks constructed of empirical data are often used as substrates for studying dynamical processes such as spreading of epidemics or information. In principle the approach is simple: we define the rules for the dynamics and run them on the network constructed of data. For example, compartmental models of epidemic spreading can be applied to model both disease and information spreading on networks [21]. Clearly, when the possible spreading pathways are restricted by real network structures, in contrast to simplified structures or fully mixed systems, the simulation results are closer to what is observed in real world [22].

To conclude the static network part of the Background section, we introduce the important concept of *percolation transition*. As the name suggests, it was originally used in mathematical physics to study how a fluid flows through a porous material. The question is whether there is a continuous open channel that connects two specified points and enables flow when the medium is organized randomly. In network setting, the analogous question is whether any two nodes are likely to be connected via a path when the links of the system are assigned randomly.

To study percolation analytically, *control* and *order* parameters are needed. When the control parameter is changed, the outcome is seen in the order parameter. Many systems go through a sudden *phase change* where the behavior of the order parameter changes from one state to another. For instance, a material may display net magnetization (order parameter) when the temperature (control parameter) drops low enough. Percolation is one example of such a phase change, where the transition happens from fragmented and disconnected phase to a connected one. The value of the control parameter where the change happens is called the *percolation threshold*.

Let us illustrate the percolation transition in networks with a simple example. As introduced, $G(N, p)$ is a random network where the existence of each link is determined independently with probability p. A percolation problem is now to study the birth of the *giant* component, *i.e.* a unique component whose size, measured in number of nodes, scales with the system size.[3] When the probability of a link p is small, all components of the system are naturally small. In the other extreme, when $p \to 1$, almost all the possible links exist and the network is a single component. If we plot the relative size of the largest component, scaled with the system size, we see a sharp transition. Also, the phase change is seen in the average size of the components other than the currently largest one. As expected, they first grow in

[3]Strictly speaking defined only in the limit $N \to \infty$ as the component that spans a finite fraction of nodes.

size but eventually join together and form the giant. The behavior is illustrated in Figure 2.2.

Analytical approaches to percolation in networks that are more complex than the random network model and closer to the ones observed in reality are also possible [23]. However, when having an empirical data set representing some network instead of some generative network model, percolation is usually approached at the other direction, *i.e.* we study how the system breaks down when nodes or links are removed or disabled, or how connectivity emerges when only a subset of the original nodes and links are considered active.

Why is it important to study percolation in empirical networks? The existence of a giant component has important consequences on the functionality of a network and especially on processes taking place on it. That is, the giant component ensures that a network is connected and thus, for instance, information can be transmitted between its nodes. If there is no giant component, the network is practically inoperable.

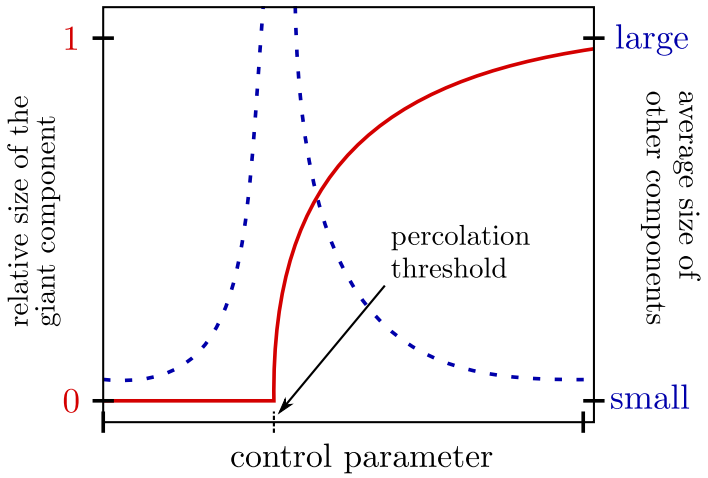


Figure 2.2: An illustration of the percolation transition and the birth of the giant component. The relative size of the giant component (red, continuous) jumps from zero at the percolation threshold and approaches one. The average size of other components than the largest (blue, dashed) diverges at the threshold.

2.2 Temporal Networks

We have already discussed processes taking place on networks. A popular approach is to define the dynamics of the process separately from the underlying network presentation: consider for example a road network as a basis for traffic simulations. However, if the underlying network is not static – the roads might have traffic lights – using a static representation either needs additional rules or results in false outcomes. This problem is approached directly by the framework of *temporal networks* (or time-dependent networks), which provides a simple representation for networks that change with time [7]. As with static networks above, we now go through the theory of temporal networks and present some notable properties of real-world temporal networks.

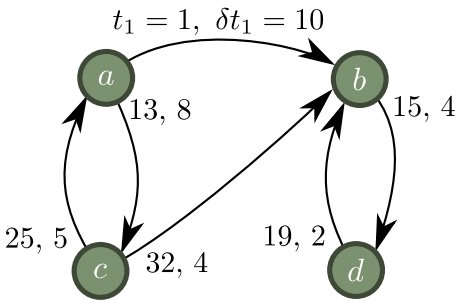
Formally, a temporal network $G_{\mathrm{T}}(V, \mathcal{E})$ is a set of *events* $\mathcal{E}$ and nodes V. Each event is a quadruple, $e \equiv (u, v, t, \delta t) \in \mathcal{E}$, where $u, v \in V$ are the nodes the event starting at time t with a duration δt connects. Thus, the events can be considered as temporal links. Since now the networks have a lifespan, we traditionally denote the difference between the ending time of the last event and the beginning time of the first event, *i.e.* the data set duration, with T. The number of events is denoted with E and the total number nodes participating in events with N. Figure 2.3 illustrates the idea of temporal networks with a common way of representing a temporal network data set as an event list.

An important concept within the theory is the *aggregated network*, which is a corresponding static version of a temporal network, constructed by assigning a link between nodes u and v if at least one event connects them. Thus, one way to conceptualize temporal networks is to consider the events as momentary activations of the underlying aggregated links. Note that the temporal (aggregated) representation allows directed and undirected definitions of the events (links). It is also common to assign weights to the links of the aggregated network based on the number of contacts or the total duration of them.

Note that this framework does not constrain the time scales of the events. With the same framework, we can represent slow changes of networks such as creating completely new links and destroying old links. Then the duration of an event would be the lifetime of the link. However, as the changes of the network structure are usually slow compared to the events occurring on the links, the underlying aggregate network is often assumed to be constant and we utilize temporal network theory on the much faster timescale of events.

The immediate question is now how are the events distributed in different real-world temporal networks and what effect this has on the function of the networks. As quantifying the properties of nodes and links is the first step in analyzing static networks, in the temporal setting we study the event sequences associated with nodes and links. The simplest metric is clearly the number of events per node, E_i, which reveals its temporal activity. If necessary, the event number can naturally be split to account for the incoming and outgoing events separately ($E_{i,\mathrm{IN}}$, $E_{i,\mathrm{OUT}}$).

(a)



(b)

t_i	δt_i	u_i	v_i
1	10	a	b
13	8	a	c
15	4	b	d
19	2	d	b
25	5	c	a
32	4	c	b

Figure 2.3: An example of a temporal network (a) and the corresponding event list representation (b). The undirected aggregated network corresponding this temporal network is seen in Fig. 2.1.

The next step is to examine how the events of a single node or link are distributed with respect to each other. For this, we study inter-event time sequences, which measure the "silent" times between consecutive events. It has been seen that in human communication, such distributions have broad tails, meaning that most events happen relatively close to each other but some have long intervals separating them. This phenomenon is know as *bursty* behavior [24]. The result is significant since it is very different from what would be expected from a random reference, which in the case of temporal networks is usually an assumption that the events are initiated by a Poisson process. This temporal clustering of events also has significant effects on the processes operating via them. For instance, it has been shown that even when the topological distances of a network are small, burstiness makes spreading processes slow in the temporal sense [25, 26]. To measure the burstiness of a node i (or a link), Reference [27] proposes a metric

$$B_i = \frac{(\sigma_\tau - \mu_\tau)}{(\sigma_\tau + \mu_\tau)} \in [-1, 1), \tag{2.5}$$

where σ_τ and μ_τ are the standard deviation and mean of the inter-event time sequence. For a maximally bursty sequence $B_i = 1$, for a Poissonian sequence $B_i = 0$ and for a completely periodic sequence $B_i = -1$.

To further study temporal properties of networks, many of the methods for static networks have their temporal counterparts. One of the most significant is the concept of *time-respecting path* [28]. As different components of a static network correspond to regions that are disconnected from each other, the addition of time may lead to temporally isolated nodes. A time-respecting path is constructed of events instead of links, and the consecutive events in a path must share a node and respect both the direction of the links and the time. That is, the next event must happen after the previous event. Whether the first event must end before the second starts is usually an application-specific decision, such as requiring that an individual can

have only one active phone call at any given time. Time-respecting paths allow us to study the reachability of each node, *i.e.* the set of other nodes that can be reached via time-respecting paths [29].

Clearly, the introduction of time for paths also gives them a *duration*. Even though a node is connected to multiple other nodes with a time-respecting path, some of the paths can be considerably faster than others, *i.e.* reach the target in a shorter amount of time. Then, similarly to the average distance between a random pair of nodes of network, we can study the average *temporal distance* (or *latency*) between them. The temporal distance between nodes u and v is defined as the shortest time it takes for u to reach v [30]. Naturally, a temporal network with short average temporal distances and high average reachability is more efficient for diffusion processes than one with long distances and low reachability.

The addition of the time dimension makes the mesoscopic examination of temporal networks significantly more difficult than the corresponding static problem. Even a proper definition of a temporal subgraph – discussed in detail in Section 3.1 – is not trivial. However, some encouraging work has been published. *Temporal motif* analysis can reveal recurrent temporal patterns of temporal networks [31, 32] and *betweennes preference* quantifies if a node prefers some time-respecting path of length two over others [33].

To this date, there are very few models of temporal networks that are capable of reproducing properties seen in empirical data. One exception is the activity driven model [34]. The key assumption of the model is that each node has an activity potential which determines the probability that it activates events in a given interval. However, the model has also many simplifications comparing to reality, such that when active, a node creates a fixed number of events that are targeted randomly to any of the other nodes.

2.2.1 Temporal Percolation

The addition of the time dimension creates a new environment for the percolation transition of a system. Topological connectivity is not enough anymore – percolation must now also respect the mutual occurrence patterns of events. Also, if we manage to define the control variable already in the time domain, that is, as an interval, then the percolation threshold would demonstrate a characteristic time scale of the system. The threshold is important since we can immediately state that any dynamical process with a shorter characteristic time scale than the threshold has no possibility of reaching the majority of the nodes of the network.

Temporal percolation has been little studied, partly because of the lack of analytical models. However, some work has been done. One approach incorporating the ideas above is presented in Reference [35] where the authors study temporal percolation in the activity driven model. They approach the problem by integrating multiple consecutive snapshots of the network up to a certain time, thus forming the control

variable as an interval, and are able to analytically predict the birth of the giant component.

In this Thesis, we approach temporal percolation with empirical contact sequences. This introduces additional aspects to be considered. First, as the underlying nodes represent real individuals, the control parameter should account for this in a reasonable way. That is, we want to have a control parameter that acts on the individual nodes and their event sequences instead of reflecting network-wide properties only. Second, nodes are typically not always active with a constant rate, *e.g.* one node may have events only in the first half of the data and another only in the second half. However both nodes must be in equal position when considering their participation in the system-level percolation transition.

Furthermore, the time domain enables us to study new concepts, such as that of a *nucleus*, a core group of nodes whose activity will ultimately form the giant component of the network. All the ideas introduced here will be considered in depth in Chapters 3 and 4 of this Thesis.

2.3 Social Networks

In the following two Sections we will show how network theory can and has been applied to the specific type of networks we utilize in this study, namely communication networks. Since we are interested in communication between humans, we start by discussing the broader concept of social networks.

Though we have already referred to social networks, we have yet to define them properly. We start with the nodes. Following the definition common in social sciences [5], the atomic building blocks of social networks – social entities – are called *actors*. Sometimes one actor node comprises multiple people: for instance, departments within a corporation, or nations of the world can be seen as one actor. However, from now on we consider one node to represent a single individual.

The definition of a link is not that straightforward. Though the underlying idea that links represent interactions between nodes still holds, the problem arises from the diversity of different possible means of interaction between humans. Importantly, as the social science term *relational tie* suggests, links can represent more abstract associations, such as kinship, friendship or mutual interests. In Table 2.1 we present possible definitions for a link in a social network, and the social relation it is based on. Clearly, there are many overlaps and ambiguities. For instance, how does one define a friend? Do all the people share the definition so that the links are reciprocal?[4] Also, as network theory enables, the links can have a weight, which now represents

[4]One of the most heartbreaking results of social network science was made by psychiatrist Jacob Moreno in the 1930s. He studied the friendships in a group elementary school students. As conventionally one might think, the boys were friends with boys and the girls with other girls, except for one boy who liked a single girl. That specific link was not bidirectional. `http://en.wikipedia.org/wiki/Network_science` (1.7.2014.).

Table 2.1: Possible definitions of a link in a social network, following References [5, 36].

type of social relation	link is based on
similarity	common spatial location
	mutual interest
	belong to the same club
	share an attribute (*e.g.* gender)
social relation	kinship (*e.g.* descendant, marriage)
	formal role (*e.g.* executive, student)
	informal status (*e.g.* friends with, knows about)
interaction	sexual contact
	face to face discussion
	collaboration (*e.g.* scientists in an article)
movement & flow	migration (*e.g.* refugees)
	transactions of resources (*e.g.* lending)
	information transfer (*e.g.* electronic messages)

the strength of a social tie. This strength can be measured via some proxies, such as basing the strength of a friendship on the number of mutual encounters (if data about them exists) or direct rankings (if executing questionnaires is possible).

Despite these underlying challenges of rigorous definitions, Borgatti *et al.* express that the power and importance of social network analysis result from the fact that it makes the individuals part of a network, and that the position within the network determines the opportunities and constraints the individual encounters [36]. When the analysis is performed at the level of individuals, the unit of interest is often an *ego network*, which studies the egos (single individuals) and their local network neighborhood.

Social network analysis has revealed many interesting phenomena arising from human interactions. For instance, it has shown the existence of *homophily*, the tendency of people to interact with others who are similar to themselves, *i.e.* share some attributes such as gender, religion or socioeconomic status [37]. A more striking result is the *weak tie* hypothesis which states that there is a positive correlation between the strength of a tie and the number of mutual acquaintances [38]. Thus, the weak links are important for the function of the network, since they can act as bridges connecting different network neighborhoods and provide one with novel information (*e.g.* hints about new jobs) or keep the network structurally intact (removing the weak ties breaks the network into multiple components [39]).

In the last few decades social network analysis has gained a lot of attention both in academia and the media. This is mostly due to the services and applications the advancements in information communication technology enables, and the popularity of their usage. To name a few, mobile phones, email, and Facebook are part of the

everyday life of many. For science, the electronic data sets most of these services store about their users and their actions are valuable since they provide details about human behavior of unprecedented magnitude and precision. A data-driven approach can then be utilized to study the social networks *per se*, and even conduct controlled experiments on *social influence*, that is, how the behavior of an individual is affected by others [6, 40–44]. Nevertheless, it's important to note that a social network constructed from a data set is not the true underlying social network but a representation of some of its facets.

Online interactions are typically recorded with time stamps. These enable us to utilize the temporal networks framework, thus making the theory of dynamical processes studied on social networks better correspond to reality. For instance, for spreading of infectious disease the order of contacts matters: one cannot get infected when in contact with someone who will only be infectious in the future [45,46]. Note that for disease spreading contacts can be direct physical contacts (*e.g.* for sexually transmitted disease) or represent shared space or location (*e.g.* germs in public transport). The setting is similar for information spreading: in order to pass the information forward, one needs to get it from someone first.

2.3.1 Communication Networks

In general, information can be transmitted in a plethora of ways. One possibility is to distribute it in one-to-many fashion, where the target is to get as large audience for piece of information as possible, and hope that the audience acquires it. To name a few, public service announcements and advertisements work this way. On the other hand, when information is distributed in one-to-one fashion between equal humans – via some medium – the contacts ultimately create an easily comprehensible communication network.

There are multiple media via which the communication between individuals can take place. The oldest and most "natural" are face-to-face discussions which can happen only if the participants simultaneously share a spatial location. Of course, today the requirement for mutual space can be relaxed with the help of technology and applications such as Skype, but the underlying nature of this fundamental communication method remains intact. Communication media can be divided into two categories depending on whether the communication requires simultaneous action from all the participants (*e.g.* calls) or whether the information transfer can be delayed and thus dependent on when the recipient acknowledges it (*e.g.* email, SMS). In Table 2.2 we present a breakdown of a few different channels of communication between individuals, and give references to network studies about them. Note that the different media differ also in whether they are strictly between two participants, as calls generally are, or whether there can be multiple recipients as in emails.

By studying communication networks we can understand both system level-phenomena, such as information diffusion in social networks, and, especially, individual-level behavior and differences between individuals. As discussed in Sections 2.2 and 2.3, the

Table 2.2: Different communication channels divided by whether they require presence in the same physical space and whether the communication requires mutual action from all the participants.

	mutual action	**delayed**
mutual space	face-to-face conversations [47]	-
via device	video conferences instant messages [14] mobile phone calls [25, 39]	traditional letters email [48–50] SMS [51, 52] Twitter [53–55]

necessary components for these studies are in place: there are multiple large data sets representing real-world communications, and there is the analytical framework of temporal networks for handling the analysis of such data.

When inferring knowledge about human communication patterns from data sets, we have to deal with two significant problems. The first is that due to privacy issues we generally do not have access to the content of the contacts; however, few studies on email communication and public channels, such as Twitter, make an exception to this. Because of the lack of content it is impossible to know the nature of the relationship between consecutive events. That is, we can't say whether there is a causal relation, *i.e.* one event is caused by another, or merely a temporal correlation. The second issue is that data sets often comprise only one communication medium at a time, even though people may communicate on multiple channels simultaneously. For instance, workplace communication is a combination of at least face-to-face, email and mobile phone conversations. To account this, there is a framework of *multiplex* networks where the nodes are connected with links from multiple classes, each class representing one communication channel [56]. Unfortunately, suitable data sets are still scarce, though some ambitious projects which aim to track the whole spectrum of communication are under way [57].

Despite these problems, analysis of communication networks has already given insight at both the network and individual levels. At the larger scale it has been seen that both the topology of the network and the occurrence times of contacts have a significant effect on dynamic processes on networks [25, 58]. At smaller scale, many interesting properties have been found. For example, there is a backbone substructure in the email network which is important for information diffusion, and that correlation between events of neighboring links occur, implying the possible causal relation between these events [49, 51]. Long-lasting (18 months) studies with mobile phone and additional survey data have revealed that the characteristic communication patterns of individuals remain constant even though the people to whom the communication is targeted change [59].

2.4 Related Work on Identifying Influential Members

Identifying important, significant or influential nodes in a network is a problem with built-in controversy: the power of any network is often that a single node usually plays a negligible role in the function of the network. Thus, removing one node or disabling it in dynamical processes does not cause differences at the large scale, because almost always alternative routes can compensate for the removed node. However, when comparing the nodes of some network between each other we can find differences in their importance to the system. To be able to perform this comparison, we need both explanatory and response variables.

For the explanatory variable we can choose any of the node-level metrics that the theory of (temporal) network analysis provides. The simplest, as discussed above, are the degree or the number of events of a node. Furthermore, there is a whole family of centrality metrics for this task. Some well-known metrics in this group are the PageRank [60] and betweenness centrality [61] which calculates the fraction of shortest paths passing through each node. Also related to these is the k-shell index of a node [62]. Nodes with high k-shell values are found closer to the topological center of the network. Also, most centrality measures have their temporal counterparts [7].

In theory, all metrics are equally valid, but for practical purposes it is better to choose metrics which can be calculated from local information, such as the degree or burstiness of a node. Also, one can criticize the meaningfulness of node-level metrics which are calculated from network-wide information especially for social networks. They reveal information about the location of node in the network but essentially assume that a node is indirectly affected by its friends' friends' friends, *etc.*

The response variable is often obtained by choosing a node with a given value for the explanatory variable and giving it a special role in some dynamical process taking place in the network. One common choice is epidemic spreading where the node is given either the role of the initial spreader or it is immunized completely. Then we average over all the nodes sharing a value of the explanatory variable and, in case of a stochastic dynamical process, over multiple runs. If the diffusion speed or prevalence is seen to vary depending on the type of nodes, they can be seen as more or less influential for the system. However it must be noted that the underlying dynamical process completely defines the outcome: if the chosen process does not represent reality, neither do the results.

This idea has been used extensively for static networks and recently the focus is shifting towards the temporal setting. The degree, k-shell index and activity of the nodes have been seen to be relatively good predictors for the importance of a node [63–65]. A few significant studies skip the bias-inducing phase of choosing the dynamical process and observe the diffusion process directly from suitable data [54, 66]. The results from empirical studies mostly agree with the simulated outcomes. Significant nodes can also be searched by considering a network as a controllable entity [67].

In this Thesis, we approach the question of the influence of a node at two scales. First, we consider influential groups of nodes, and then influential individuals.

As will be shown, the nucleus of a temporal network, *i.e.* the set of nodes that will give rise to a giant component, is important for the system. Thus, the nodes who are responsible for creating the nucleus are more influential than a random group of nodes in any process taking place on the network. We will study whether it is possible, given just properties of a node that can be calculated from its local information, to predict whether it is part of the nucleus.

At the individual level, we treat the term "influential" in a more literal way. That is, we study the role of nodes within their local neighborhoods, and see whether some nodes are good for continuing information spreading, and whether some nodes create more action than others. These phenomena are reflected against the same node properties used already in the influential group studies.

3 Methods

In this Chapter we will go through the methods used in this study. First, we will discuss how a temporal network can be split into temporal subgraphs in a meaningful way which enables observing the percolation transition. Then, we will define a method for labeling the events within each subgraph based on their role. This allows to see what kind of role each node has in the given subgraph and to study the node's level of influence. As discussed above, comparing observed results to a reference is always important. In Section 3.2 we introduce the reference model used in this Thesis and the general convention for creating other reference models. This Chapter is concluded by presenting the used data sets and their basic properties.

3.1 Temporal Subgraphs

To be able to study temporal networks not just as an complete entity but at a smaller scale which is suitable for studies of individual nodes, we first need to break the full network into smaller parts, to *temporal subgraphs*.

Basically, a temporal subgraph (TSG) $G_{\mathrm{T}}(V', \mathcal{E}')$ is any subset of the nodes and events from the original temporal network, $V' \subseteq V$ and $\mathcal{E}' \subseteq \mathcal{E}$. However, taking an arbitrary subset of both results in subgraphs which are unusable for inferring information. In order to give the subgraphs a sensible interpretation, the rules controlling their creation must embody the features of the underlying temporal communication network. For constructing such TSGs, we follow the definitions of Reference [32].

First, we define two events to be Δt-*adjacent* if two events happen within a time span of Δt and share at least one common node. Note that if the events have duration other than zero, then the time difference between the ending time of the first event and the starting time of the second event cannot exceed Δt. Also note that the direction of events does not matter. With the help of the adjacency we can define Δt-*connectivity* of any two events: events e_i and e_j are Δt-connected if and only if a sequence of Δt-adjacent events connects the two. Then, with the Δt-connectivity we can give an unambiguous definition of a temporal subgraph: a temporal subgraph is a maximal set of Δt-connected events, that is, a set where no more Δt-connected events can be added.[5]

The definition is deterministic in the sense that we always end up with the same set of subgraphs with a specific Δt, independently of the order in which the events are encountered. Since Δt-adjacency is independent of the direction of the events, also non-causal event patterns, such as an outgoing event before incoming event are included. Note that each event can belong to one TSG only, whereas nodes can participate in multiple TSGs (and also appear multiple times in the event sequence

[5]Note that in Ref. [32] the definition for the temporal subgraph is does not require the maximal set of events but only that all the consecutive events are included.

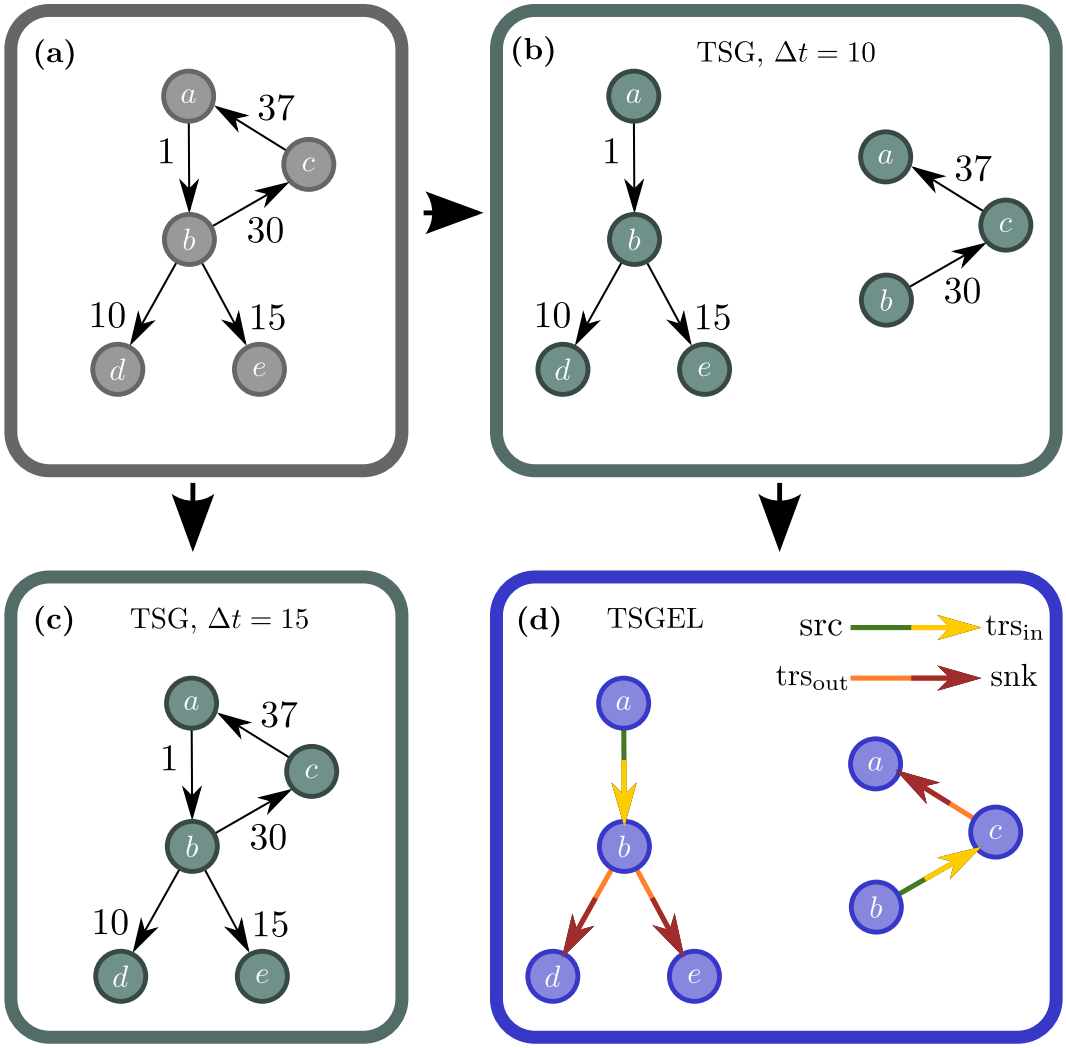


Figure 3.1: A schematic for constructing the temporal subgraphs and the event labeling rules. The temporal network with duration $\delta = 0$ events of panel (a) is broken into subgraphs. In panel (b) using $\Delta t = 10$ results in two subgraphs. With $\Delta = 15$ (c) all the events are in a single subgraph. The labeling of the events of the subgraphs in panel (c) is seen in panel (d). Note that the event starting at 15 gets a transmitter label to its root, since it is Δt-connected to an incoming event via another previous outgoing event.

of one TSG). One way to think of the role of the parameter Δt is to consider it as a waiting time counter for a node that is activated after an event where the node participates ends. If an event occurs to the same node before the counter reaches zero, the event is added to the existing TSG.

This definition of a temporal subgraph avoids a few common problems that exist in other methods for breaking down a temporal network into smaller pieces. For example, a common method of studying subgraphs defined by predefined time-slices [68,69] (for example, construct a network separately for each day of the data) loses the possible differences at shorter time scales than the slice and introduces arbitrary boundaries where the subgraphs are split. The method used in this Thesis

is data-driven in the sense that the subgraphs emerge from the true activity of the individuals and thus it enables comparison between different subgraphs. Another possible approach would be to study causal time-respecting paths which are created by Δt-connected events, where the causality means that the direction of the events must enable flow. However, the causality condition is very restrictive and results in small temporal subgraphs when Δt is small. When Δt is large, flow paths start to split and include parts of themselves, leading to combinatorial problems.

One benefit of the TSG method is that the computational complexity of calculating the TSGs is $\mathcal{O}(E)$ if the events are properly sorted. The algorithm is also very simple. Basically, we choose any event that is not assigned to a TSG as a seed of a new TSG, go both forward and backward in time to find all the Δt-adjacent events, and add them to the TSG. This is repeated recursively until no new events are found. At the end of one iteration we have found one TSG and can move to the seed event of the next one, if there are unassigned events left. A pseudocode implementation of the algorithm is presented in Algorithm 3.1 and an illustration of the method is displayed in Figure 3.1.

3.1.1 Temporal Subgraphs with Event Labeling

Temporal subgraphs provide a tool for splitting temporal networks into smaller pieces and form the basis of studies at the node level. If we in addition label the events according to their role within a given TSG and use these labels to give scores to the nodes, we can then utilize this information to analyze the role of the nodes and their influence. To the best of authors' knowledge, the proposed method of temporal subgraphs with event labeling (TSGEL) method has not been introduced elsewhere.

The fundamental idea behind the labels is the source/transmitter/sink -paradigm where nodes initiate information flows, act as relays further transmitting information, or receive but do not further transmit information. Such flows are based on causality in the direction and order of events. As the events are directed, we consider the roots and tips of the events separately and assign them a label according to other previous or future events, if any. The rules to label event e of a given TSG are:

	label	**rule**
root	source	no Δt-connected incoming event to the root node before e
	transmitter	otherwise
tip	transmitter	at least one Δt-adjacent outgoing event in the tip node after e
	sink	otherwise

Thus we have four labels altogether. Note that the root uses Δt-connectivity whereas the tip uses Δt-adjacency. This is because an event root can be Δt-connected to a

previous incoming event via other outgoing events, and then it is intuitive to give it a transmitter label as the other outgoing event root gets a transmitter label. In other words, if we assume that an incoming event e_i and a later outgoing event e_j of node w are adjacent with some Δt, we must assume that an even later outgoing event e_k of node w is also related to the incoming event e_i, if e_k is Δt-adjacent to e_j. In the tips of the events the situation is clearer and we need to only look for the Δt-adjacent outgoing event in the tip node. Panel (d) of Figure 3.1 shows an example of the labeling method.

The labeling of the events can be done simultaneously when calculating the subgraphs. Note that the method also works when the events have zero duration, and multiple events can start or end for a given node at the same instant. However, the underlying idea of the labels, that one event causes others, is lost since immediate events clearly can't represent causal human behavior. Also note that the method does not allow temporally overlapping events.

Algorithm 3.1 Algorithm to find the temporal subgraphs of a temporal network G_{T} with a given Δt.

Require: $\mathcal{E}$ is a set of events of a temporal network G_{T} that do not overlap for any of the nodes, and $\mathcal{E}.i$ is the subset of events where one of the participants is node i. Each $e \in \mathcal{E}$ has fields $e.t$, $e.\delta t$, $e.u$, $e.v$ denoting the starting time, duration, initiator and receiver of the event, respectively. In addition, the field $e.assigned \in \{\texttt{True},\texttt{False}\}$ indicates whether the given event is already assigned to a TSG and is initially `False` for all events.

function TSGFinder($\mathcal{E}, \Delta t$)
    TSGs $\leftarrow \emptyset$
    for e in $\mathcal{E}$ **do**
        if not $e.assigned$ **then** ▷ if event is not in TSG
            oneTSG $\leftarrow \emptyset$
            $e.assigned \leftarrow$ `True`
            eQ push e ▷ push e to the empty event queue
            while eQ $\neq \emptyset$ **do**
                e_{curr} pop eQ
                oneTSG $\leftarrow$ oneTSG $\cup \{e_{curr}\}$
                dtAdjacent($\mathcal{E}, \Delta t, e_{curr}, eQ$)

            TSGs $\leftarrow$ TSGs $\cup$ oneTSG ▷ add the found TSG to TSGs container
    return TSGs

function dtAdjacent($\mathcal{E}, \Delta t, e, eQ$) ▷ push Δt-adjacent events of e to eQ
    for e_c in $\{\mathcal{E}.(e.u), \mathcal{E}.(e.v)\}$ **do** ▷ candidate Δt-adjacent events for e
        if not $e_c.assigned$ **then**
            if $e_c.t \geq e.t + e.\delta t$ and $e_c.t \leq e.t + e.\delta t + \Delta t$ **then** ▷ e_c after e
                $e_c.assigned \leftarrow$ `True`
                eQ push e_c
            else if $e_c.t + e_c.\delta t \leq e.t$ and $e_c.t + e_c.\delta t + \Delta t \geq e.t$ **then**
                $e_c.assigned \leftarrow$ `True`
                eQ push e_c

3.2 Random Time Shuffle Reference

As already discussed, we need something to compare the results with. The standard convention in science would be to organize a controlled experiment where we are able to divide the participants to treatment and control groups. Clearly this is not possible, since we are not able – or ethically allowed – to intervene in human activities. Thus, we need to construct a reference from the data we already have.

With empirical temporal networks, the established practice in creating a reference is to shuffle the events of the original data with different rules [7]. This way we can control which event correlations we want to preserve and which destroy. Then, the difference between the results with the shuffled data and the original must originate from the removed correlations. In principle, the shuffling methods are employed when we are interested in phenomena at the level of the whole network and time scales of the order of the data duration T. For individual nodes or links the alterations are either trivial or difficult to interpret.

In this Thesis, we use the random time shuffling (RTS) of the events, where the starting times and durations of the events are randomly redistributed. Note that the starting time and the duration are always switched together, thus keeping the original events intact. The RTS method destroys correlations of event timings both within and between links but preserves the underlying aggregated network, number of events of nodes and links, and the circadian patterns (to be introduced shortly). Since we need to preserve the validity of the data with non-zero event durations, *i.e.* the events of a node must not overlap, the shuffling is executed with the Markov Chain Monte Carlo method. At each step, two events are chosen uniformly at random and their switching is accepted if it does not create overlapping events. The method is halted after $5 \times E$ successful switches.

3.3 Data Sets

In this Thesis we use three empirical data sets representing human communication via different channels: mobile phone calls, mobile phone SMS messages, and emails. The first two are constructed from the same mobile phone data set. Next we present the data sets, their pre-processing and their basic statistical properties.

3.3.1 Mobile Phone Data

The mobile phone data comes from the billing system of an European carrier with a market share of $\sim 20\%$ of the population within that country.[6] We have access to data for one entire year (2007) but will use shorter data periods since these are computationally more convenient. Both mobile phone calls and SMS messages are included in the same data set. Essentially, the data is a list of customer IDs (a

[6]We thank A.-L. Barabási of Northeastern University for the mobile phone data.

hashed phone number) and their time-stamped connections via either calls or SMS messages. The hashed ID's guarantee anonymity; the data cannot be linked to persons. Note that we naturally do not have any information on the content of calls and messages, and hence no definite conclusions on the possible causality of any event sequences can be drawn. For instance, if B calls C after talking with A, we cannot know if the latter call was triggered by the former.

In addition to the call and SMS events, we have access to some demographic information for the users who are customers of that specific carrier. We call these users the *company users*. The raw data also contains the connections to phone numbers which do not belong to customers of this company (*i.e.* the *non-company users*), but we will shortly justify why they are discarded from the analysis completely. The demographic information for the company users contains for instance age and gender of the customer, activation and disconnection dates of the contract and whether the contract has prepaid or postpaid billing. Unfortunately, many customers have incomplete information and partly due this, the demographic information is not used in this Thesis.

Though the calls and SMS are mixed in the original data, as they are also mixed as communication channels, we consider them as a separate temporal networks. This separation is possible since calling and messaging are mostly used for different types of communication. The different nature of SMS messages and calls is seen *e.g.* in their correlations and timings – messages typically appear as repeated "ping-pong" event strings between two customers, whereas call patterns are more diverse [51].

The raw data goes through multiple pre-processing steps in order to filter anomalies from the data and, especially, to ensure that the data is as good representation of the underlying social network as possible. The steps, and the reason why they are executed, are:

1. Remove corrupted events *i.e.* events that have erroneous information in either the duration or the cost columns of raw billing data. Examples are negative cost value or zero duration value for a call.
2. Sort the events by their start time.
3. Remove all the events where at least one participant is a non-company user. Even though this makes the network more sparse, this step is essential since we can not be sure about the reliability of the non-company users and their events (the number of calls made by non-company users varies abnormally, indicating that the billing system records these events differently than company events).
4. Remove all events that do not take place on a mutual link. A mutual link is a link where at least one event occurs in both directions. This step helps to reduce anomalies caused by *e.g.* marketing calls and call centers and therefore the remaining events correspond better to real social interactions.
5. Check that there are no overlapping calls for any of the nodes. An event starting at the same second than the previous ends is accepted. Since the

SMS messages have duration zero, they cannot overlap.

In our analysis we use two different sets of call data, where the duration of the shorter one is one month and the duration of the longer one is six months. The first month of the six-month data is the same as the one-month data. The order of magnitude of the number of nodes is 10^6 for both mobile phone networks.

3.3.2 Email Data

The email data represents the email communication network extracted from the logs of a university's email server [50]. This data is received and we use it as it is – it is known that the pre-processing steps applied at the data source include removing other than intra-institute messages and certain mass mailers. The data spans a period of 82 days and it represents email communication between 2997 individuals. The real occurrence times of the messages have been concealed by offsetting the time stamps of the events. However, the offset is the same for all events, keeping the intervals between events and circadian patterns intact.

3.3.3 Basic Properties of the Communication Networks

Though all the three data sets represent human communication via electronic devices, they have some underlying differences. Mobile phone calls require activity from both of participants, *i.e.* the call must be picked up for it to appear in the records, whereas for SMS and email only the sender is required to be active. Also, the messages are not necessarily read immediately upon reception, if they are read at all. However, we make the assumption that the information in the message is available to the user immediately as the message is sent, and thus consider them as events with duration zero.

The structure of the underlying aggregated network also varies between the data sets. Table 3.1 shows some basic statistics calculated for the undirected version of the aggregated network alongside with temporal features. We see for example that the call network is denser than the SMS network, visible in the larger values of mean degree $\bar{k}_i$ and mean clustering coefficient $\bar{C}_i$. The email network is clearly the smallest and has the highest average degree, but this is partly the result of constricting the data collection only to the dense network of intra-institute messaging.

Table 3.1: Basic statistical properties of the temporal networks constructed from the different data sets. The topological network properties are calculated for the aggregated network with undirected links. The values are from left to right: number of nodes in the network N, number of events E, number of links m, average degree of the nodes $\bar{k}_i$, average clustering coefficient of the nodes $\bar{C}_i$ (nodes with $k_i < 2$ omitted), average number of events per node $\bar{E}_i$, average number of events per day, the relative size or the largest connected component (LCC), time interval spanned by data, *i.e.* data duration T, data resolution, and event duration (- denotes an arbitrary duration).

	N	E	m	$\bar{k}_i$	$\bar{C}_i$	$\bar{E}_i$	E per day	$\|\text{LCC}\|/N$	T	resolution	e duration
Call	4.96×10^6	88×10^6	8.6×10^6	3.5	0.30	17.8	28×10^5	0.88	31 d	1 s	-
Call 6M	6.28×10^6	552×10^6	18×10^6	5.6	0.28	87.9	30×10^5	0.96	181 d	1 s	-
SMS	3.07×10^6	34×10^6	4.0×10^6	2.6	0.13	11.0	11×10^5	0.79	31 d	1 s	0
Email	2997	20×10^4	2.2×10^4	14.5	0.28	67.6	2472	0.999	82 d	1 s	0

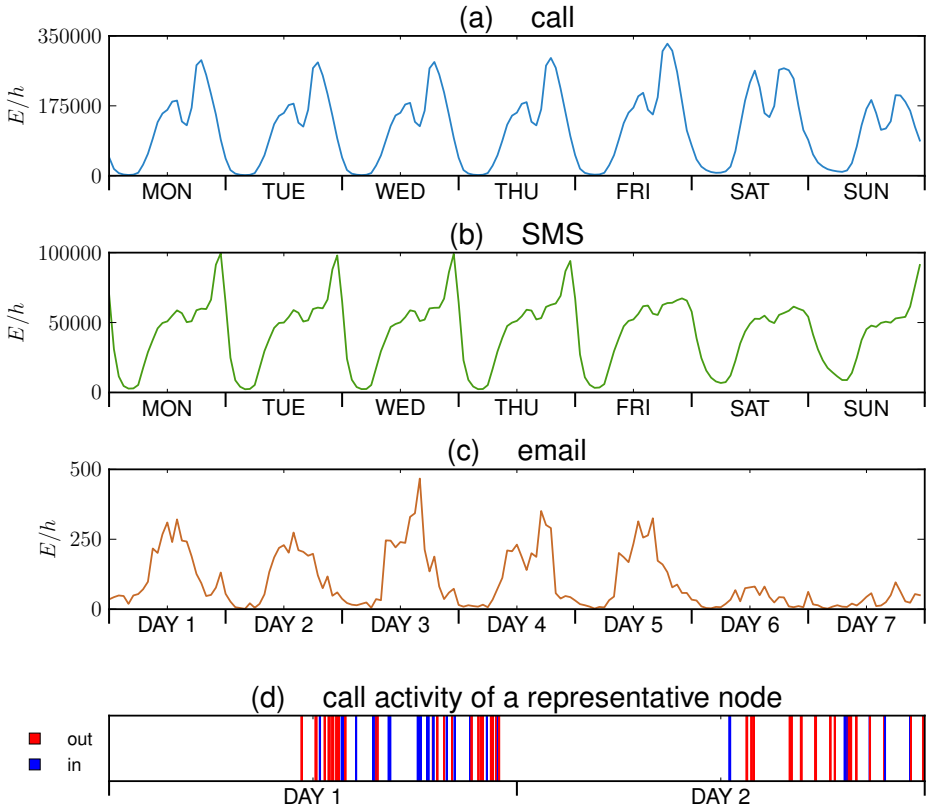


Figure 3.2: Circadian and weekly patterns of the temporal networks and an example of the bursty behavior of an example node. In panels (a), (b) and (c) we see the hourly number of events for call, SMS and email data, respectively. The circadian and weekly patterns are clearly visible. For mobile phone data the week begins on Mon 8th Jan 2007, 00:00 (local time). For email data the day borders are matched to the presumed weekly pattern. In panel (d) we see the bursty behavior in outgoing (red) and incoming (blue) calls of a single node during an interval of two days.

In the temporal domain the simplest metrics are the averages of event counts per node and per day. The email network has the most events per node and the SMS network the fewest. In the call data, the daily number of events is approximately the same for the 1-month and 6-month data sets, which is a sign of the stability of the data set. It is also natural that the number of events per node does not grow as fast as the time span covered by the data, since nodes contribute to the total event count only when they are active.

The circadian pattern, that is, low activity during the night and high activity during the day, and differences between different weekdays are seen in Figure 3.2. In the mobile phone data we see a daily bimodal behavior in the hourly number of events which is most likely caused by the transition from office calls during the day to private calls during the evening. Email communication slows down for the weekend as it is mostly used for work. We also show the bursty nature of the call activity of one single individual in panel (d) of the figure.

4 Results

This Chapter is divided in five Sections where we report the results of the analysis, proceeding from system-level to node level. First, in Section 4.1 we see how the temporal networks split into temporal subgraphs as the parameter Δt that defines temporal adjacency of events belonging to the same subgraph is varied, and study temporal percolation in the empirical networks. This is followed by a separate discussion Section. Next, in Section 4.3 we study if there are differences in the local topological and temporal properties of the nodes who are responsible for the system-wide percolation transition seen in the previous Section. In Section 4.4 we use the TSGEL method for addressing whether some individual nodes are more influential than others. The Chapter is concluded with a discussion on influential nodes.

4.1 Temporal Percolation in Communication Networks

In Figures 4.1, 4.2 and 4.3 we see how the temporal networks break into TSGs when the parameter Δt is varied for the call, SMS and email data sets, respectively. We refer to these figures in the following five subsections where the temporal percolation transition is discussed.

4.1.1 Distribution of TSGs

In panels (a) and (b) we show the size distribution of the TSGs, measured in either nodes or events, for various values of the parameter Δt. The corresponding outcomes for the RTS reference model are shown with dashed lines. These plots are similar to those shown in Reference [31]. For all the data sets we see four common features.

The first is that the distributions are broad and widen as the parameter Δt is increased, reflecting the natural growth of the subgraphs when Δt is increased.

The second important feature is that the RTS reference yields smaller temporal subgraphs; a clear indication that there are temporal correlations in the original data that enhance temporal subgraph formation. For instance, in call data with $\Delta t = 300$ s, 5.4% of the subgraphs in original data have at least three nodes, whereas in the RTS reference the corresponding fraction is 1.6%.

Third, and related to the second point, the shuffled outcomes are closer to the original when Δt is larger. This indicates the time scale where the shuffling has an effect: if the outcomes are similar, increasing Δt compensates for the destroyed correlations between events.

Fourth, as the parameter Δt is further increased, the main difference in the distributions is in their tails, *i.e.* the largest temporal subgraphs are of different size. This implies a transition in the system-wide behavior to a regime where the only considerable change as a function of the parameter Δt is in the size of the largest

temporal subgraph. This is a clear indication that a temporal percolation is taking place: smaller temporal subgraphs merge into the largest TSG. Next, we inspect how this transition happens in more detail.

4.1.2 Rise of the Giant TSG

In panels (c) and (d) we study the relative size of the largest temporal subgraph $\mathrm{TSG}_{\max}$ with a given Δt, measured both in nodes and in events. The normalizing quantity is the total number of nodes N or events E in the network. On the right-hand-side vertical axis we superimpose the average size of all other TSGs, that is, excluding $\mathrm{TSG}_{\max}$. In all the data sets we clearly see a phase change where the size of the $\mathrm{TSG}_{\max}$ starts to encompass a large fraction of the nodes and the average size of all other TSGs has a local maximum. Note that for the email data, we see the most textbooklike phase transition in the sense that the average TSG size returns to small values after the peak. This is because the underlying network is essentially only one component and eventually almost all the nodes join it, whereas in the call and SMS networks the components other than $\mathrm{TSG}_{\max}$ can grow independently (see Table 3.1).

Clearly, we are seeing a percolation-like transition in the networks where a giant temporal subgraph is born. Before we can answer the important question where the percolation exactly happens, *i.e.* where the percolation threshold exactly is, we need to address the differences between static and temporal percolation. In the static case we are interested in the connectivity of the network and study the emergence of the giant component, *i.e.* the unique largest component, as a function of some control parameter. For static networks, behavior of various quantities around the percolation threshold is well-known. Although this theory holds strictly speaking only for infinite networks (or ensembles), for any network, the percolation threshold can be approximated as the point where a large component emerges and various quantities – such as the average size of other components – diverges. However, for temporal networks this situation is much more difficult. Consider the giant component – it is not enough that it becomes large, it should also be long-lived.

4.1.3 Lifetime of the Largest TSG

Thus, in order to understand when temporal percolation takes place, we focus on the lifetime of TSGs. The lifetime of a temporal subgraph is the time difference between the beginning of the first event and the end of the last event. Clearly, when Δt is increased, the lifetime of subgraphs increases as well, since the requirement of temporal proximity of adjacent events is relaxed. The parameter Δt in itself offers us a reference: if the lifetime of a TSG surpasses the reference, it is a sign that adjacent events within that specific TSG happen close enough in time so that they keep the TSG alive. Thus, the lifetime is due to event correlations and not merely due to the parameter. Note that with small values of Δt also the durations of the

events can result in lifetimes exceeding the reference. Next we study the lifetime of the largest subgraph $\mathrm{TSG}_{\max}$, size measured in nodes, which should be sufficient to reveal possible system-wide phenomena.

In panel (e) of the figures we see the relative lifetime of the $\mathrm{TSG}_{\max}$, where normalizing is done with the data time span T. Also, the reference $\Delta t/T$ is shown. In all data sets we see that the lifetime of the largest subgraph exceeds the reference at approximately one hour. Yet, the remarkable outcome is the stepwise behavior of the relative lifetime. The system goes from lifetimes barely exceeding the reference to a state where the $\mathrm{TSG}_{\max}$ is present for the entire time span of the data. This is a clear indication of a temporal connectivity emerging: when Δt is increased above this threshold the subgraph is kept alive by the event correlations and the parameter becomes irrelevant.

The lifetime of the largest subgraph is a significant factor when defining the percolation threshold. As we have not restricted the nodes in the networks based on their activity, a node can be active only during the first half of the data. However, when discussing system-wide phenomena we need to confirm that all the nodes have at least in theory a chance of joining the giant component. This can be ensured by requiring that the lifetime of the $\mathrm{TSG}_{\max} \sim T$ which results in the equal treatment of the nodes irrespective of their activity periods.

4.1.4 Uniqueness of the Largest TSG

Although the substantial lifetime of the TSG is a necessary condition, it is not a sufficient condition for defining the percolation threshold and the birth of the giant component. It is possible that there are multiple components which reach large size and long lifetimes and the $\mathrm{TSG}_{\max}$ alternates between these. Thus, for determining the uniqueness of the giant component, it is important to verify that the set of nodes assigned to the largest component does not change when Δt is increased. For this, we define a stability metric

$$S(\Delta t_i) = \frac{|\mathrm{TSG}_{\max}(\Delta t_i) \cap \mathrm{TSG}_{\max}(\Delta t_{i+1}) \cap \ldots \cap \mathrm{TSG}_{\max}(\Delta t_{i_{\max}})|}{|\mathrm{TSG}_{\max}(\Delta t_i)|}, \tag{4.1}$$

where the numerator is the number of nodes who are in the $\mathrm{TSG}_{\max}$ with Δt_i and also present in the largest component with all larger Δt values, and the denominator is the size of $\mathrm{TSG}_{\max}$ at Δt_i measured in nodes. When this metric stabilizes to 1, the core group of nodes that form the unique giant component has been found. When Δt is increased, new nodes join this core. In the following, we will refer to this core group as the *nucleus* of the giant component.

The behavior of Equation (4.1) for each of the networks is seen in panel (f) of the figures. In the call and SMS networks, we see a sharp transition from $S = 0$ to $S = 1$, whereas in the email data S is rather high (~ 0.8) already for the smallest Δt and reaches $S = 1$ less sharply.

It is also important to relate the value of Δt with the different indicators of the phase transition. In the call data, all the indicators point towards critical Δt's that are close to each other. In the SMS network, the core group of nodes stabilizes at much lower Δt than the other signs of phase transition take place. This is most likely due to the "ping-pong" nature of SMS messaging: nodes who differ from this behavior and send messages to multiple recipients within a short interval stand out as a group. In the email data, the special features of the data set stand out. Because the network is effectively just the LCC, the phase transition is seen in terms of nodes earlier than in events. Also, because the activity in the email network goes down significantly during the weekend, it results in the death of temporal subgraphs, late stabilization of the relative lifetime, and oscillations in the stability measure.

4.1.5 Temporal Percolation Threshold

The final task in percolation studies is fixing the temporal percolation threshold. Thus we want to specify the value of parameter Δt where the nucleus of the giant component emerges, and denote this with $\Delta t = \Delta t_c$. As stated, in the temporal setting the size of the largest subgraph is not enough, we also need to observe the lifetime and the specific nodes forming the nucleus of the giant component. Hereby, we estimate Δt_c based on when the relative lifetime of the $\mathrm{TSG}_{\max}$ and the stability of the nodes (Equation (4.1)) have stabilized to one. The values of the percolation threshold, alongside with the corresponding relative size of the $\mathrm{TSG}_{\max}$ are reported in Table 4.1.

Table 4.1: Percolation threshold Δt_c and the corresponding relative size of the $\mathrm{TSG}_{\max}$ for the call, SMS, and email networks, respectively.

Data set	Δt_c	$\lvert\mathrm{TSG}_{\max}(\Delta t_c)\rvert/N$
Call	18000 s	0.10
SMS	18000 s	0.02
Email	86400 s	0.98

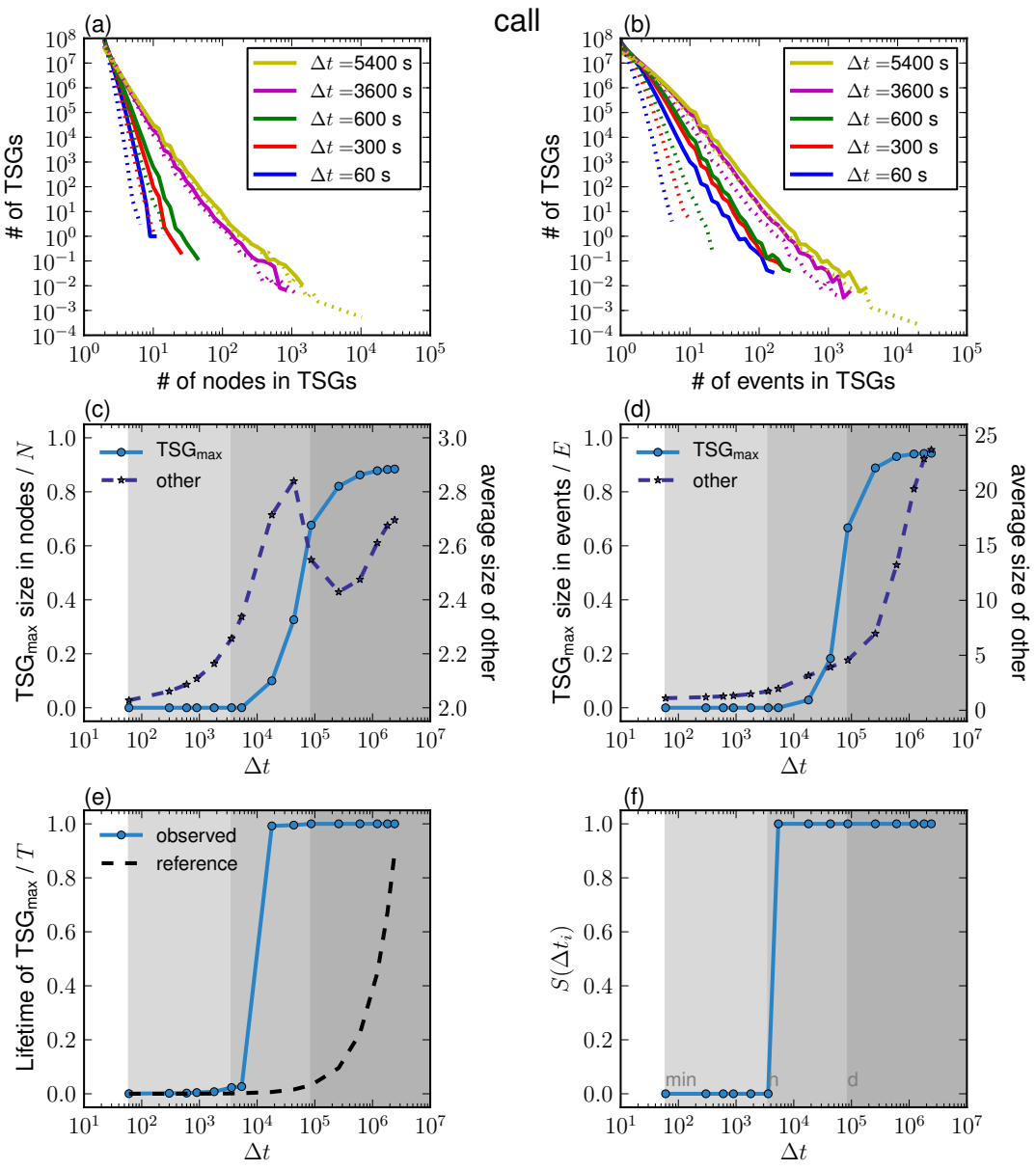


Figure 4.1: Temporal percolation in the call network and the birth of the unique giant component.
(a) & (b) Histograms of the number of TSGs of given size measured in nodes or events. The dashed lines depict results for the RTS reference. The event correlations present in the original sequence form larger TSGs compared to the reference but the difference vanishes with larger values of Δt.
(c) & (d) The relative size of the largest TSG and the average size of all the other TSGs measured in both number of nodes and events. The system goes through a phase transition where a giant TSG emerges.
(e) & (f) The lifetime of the largest TSG and the stability of nodes within the largest TSG. Both measures reflect the stability and uniqueness of the giant component and are important when determining the percolation threshold and the nucleus of the giant component. The shading in panels (c)-(f) emphasizes intervals of one minute, one hour and one day.

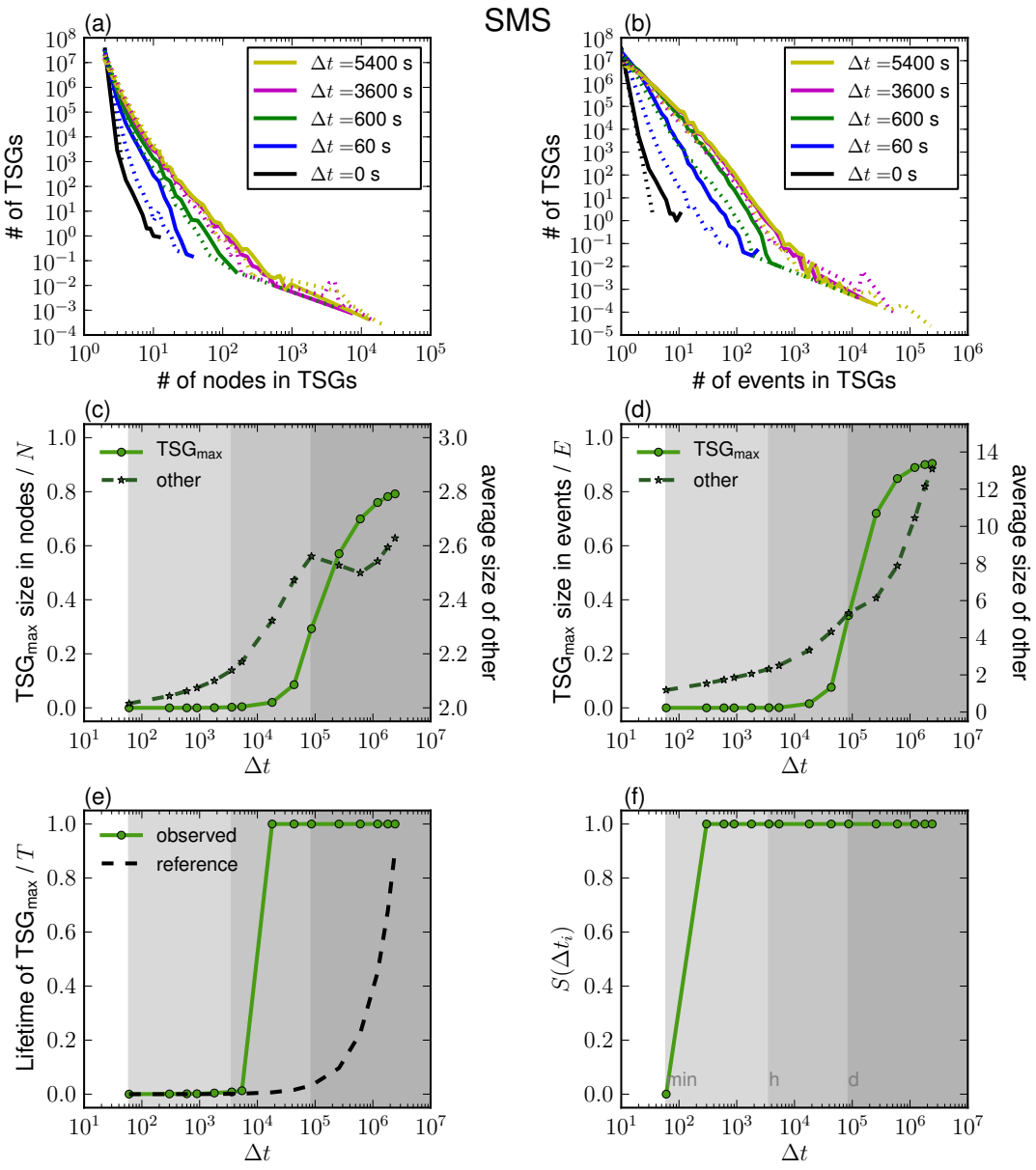


Figure 4.2: Temporal percolation in the SMS network and the birth of the unique giant component. In (a) & (b) the dashed lines depict results for the RTS reference and the shading in panels (c)-(f) emphasizes intervals of one minute, one hour and one day. System-wide behavior resembles qualitatively that seen with call data and explained in Figure 4.1 with one distinction: in panel (f), the set of nodes present in the largest TSG stabilizes much earlier. This difference is most likely due to the different nature of SMS communication compared to call communication. The group of nodes who send messages to multiple recipients within a short interval stand out from the rest, who typically exchange messages with a single recipient only.

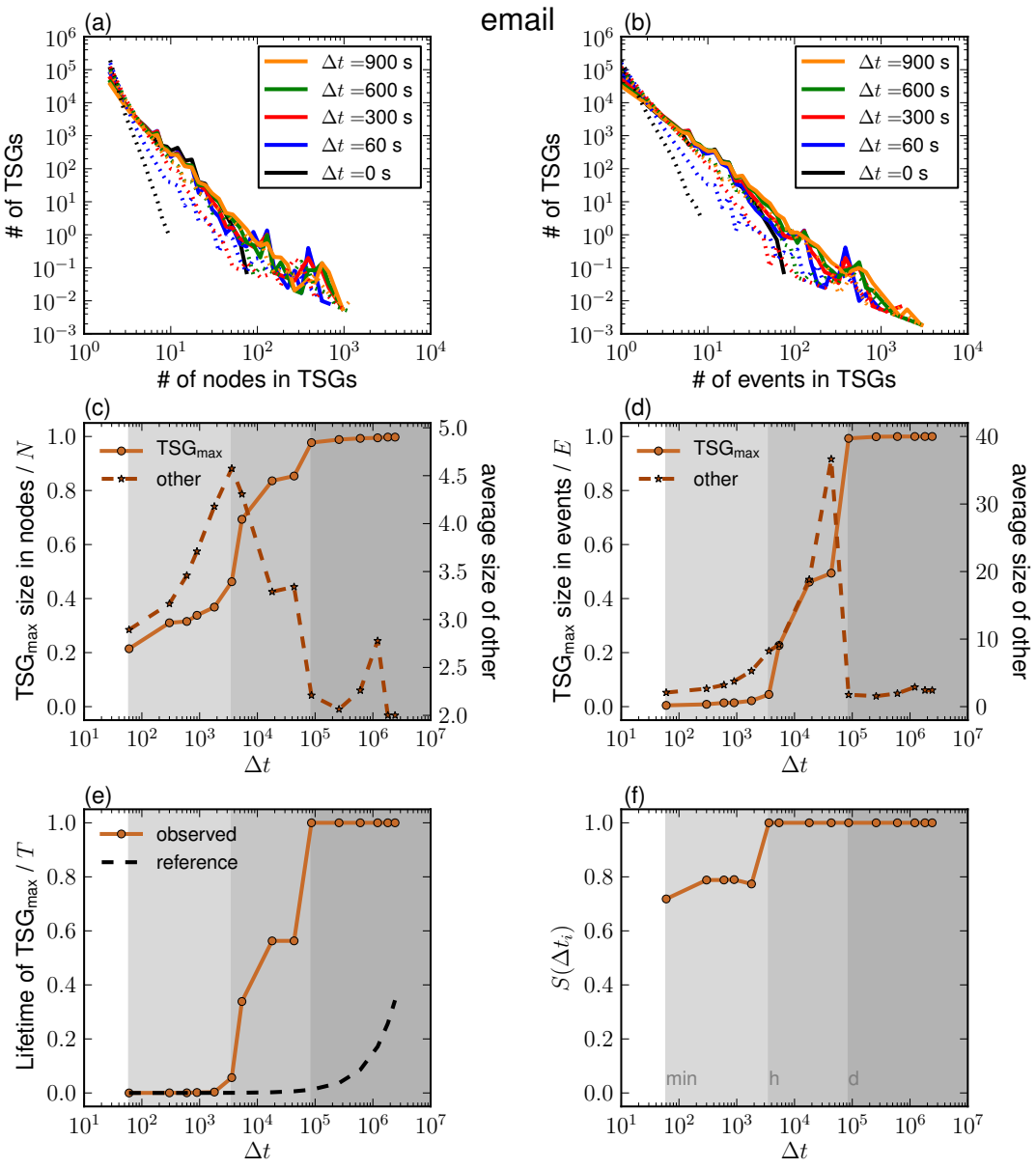


Figure 4.3: Temporal percolation in the email network and the birth of the unique giant component. In (a) & (b) the dashed lines depict results for the RTS reference and the shading in panels (c)-(f) emphasizes intervals of one minute, one hour and one day. System-wide behavior resembles qualitatively that seen with call data and explained in Figure 4.1 with two distinctions. The first is that in the phase transition plots of panels (c) and (d), the curves of the average size of other TSGs measured in nodes and events have only one sizeable peak, after which they remain at low values. This is because in the email network 99.9% of the nodes are in the LCC and the average size of other components consists of only the few remaining nodes and the TSG formed by them can not grow. The second distinction is that the relative lifetime of the $\mathrm{TSG}_{\max}$ reaches one only with quite large parameter values. This is due to the low number of messages during the weekend which causes the death of the subgraphs.

4.2 Discussion on Temporal Percolation

It is important to note that the approach to temporal percolation presented above is not the only definitive one. Clearly, the underlying framework of temporal subgraphs determines percolation behavior, and as we already discussed in Section 3.1, it is also possible to create subgraphs with stricter or more relaxed rules. If we tighten the rules and require purely causal paths between the nodes (*i.e.* in addition to the Δt-adjacency, we consider the direction of the events) we effectively study the longest time-respecting path. However, this would result in percolation thresholds with significantly higher Δt. Observations in Reference [30] support this claim. On the other hand, we can relax the Δt-adjacency requirement by assigning a common timer to all those nodes who are already assigned to a subgraph. This means that an event which takes place within Δt of the last event of any of the nodes in the TSG would be included, irrespectively of the previous or future events of that node. Clearly, this would allow for percolation at lower values of Δt, but also makes interpreting the results more difficult. For instance, do completely separated events have a meaning in a communication network? Thus, the TSG method we applied can be seen as an intermediate form of these two modifications and most suitable for human communication networks.

The characteristics of the underlying networks also need to be considered. For instance, is the requirement of the substantial lifetime of the giant component reasonable if the event activity varies significantly? Since circadian patterns are typically strong in networks representing human communication and thus the activity is very low during the night, one could argue that studying the maximal temporal subgraph only during the hours of the daytime would be enough. This is – indeed – a valid question for further work. Also, studying how the active nodes change *within* the $\mathrm{TSG}_{\max}$ is interesting. That is, intuitively most nodes are active during the day, but the nightly activity of some nodes keeps the giant component alive. However, these are more application-specific questions and do not diminish the importance of the concept of the TSG lifetime when studying the temporal percolation transition at the system level.

It is also worth to noting that the definition of the TSG lifetime is independent of the method chosen for constructing the subgraphs. Thus, it can be utilized with any other definition of a subgraph. The only restriction is that the lifetime of each subgraph must be driven by the events. Yet, as the observation of the large Δt_c for the email network indicates, the concept of lifetime should always be used with care.

The TSG lifetime should also be stable with data time span T, if the underlying network does not vary too much and T is sufficiently larger than the threshold Δt_c. Observations when comparing the thresholds between the one-month and six-month call data support this claim: the percolation threshold is exactly the same. However, the relative size of the $\mathrm{TSG}_{\max}$ with the longer data is larger, now $\sim 30\%$. This increase is expected as a longer time enables more nodes to have events that will

join them to the $\mathrm{TSG}_{\mathrm{max}}$.

Finally, what does the percolation threshold of 5 hours for the call data actually represent? It is a kind of a characteristic time scale that combines features from both node and network scales. That is, at the node level the adjacent events of single node must be temporally close enough so that the TSG stays alive and spreads. At the network level, sufficiently many nodes behaving in this fashion must be neighbors of each other so that the TSG becomes unique and large enough. The most meaningful interpretation for the percolation threshold is when comparing it with epidemic spreading where the nodes can recover (see the SIR model in *e.g.* [8]). Any process with a characteristic time scale for node activity attenuation that is less than Δt_c cannot reach a significant proportion of a temporal network's nodes or remain alive for the whole time span of the network.

4.3 Influential Groups of Nodes

After identifying the birth of the giant temporal subgraph in temporal networks, the next immediate question is that whether the nodes responsible for its emergence – the nucleus of the giant component – are more influential for the network than a random group of nodes of similar size? We answer this question in the Section 4.3.1. After that, in Section 4.3.2, we will study whether this group can be separated from the rest on the basis of the local properties of its nodes.

4.3.1 Significance of the Nucleus for the Network

As discussed, we define the nucleus as the set of nodes forming the giant component of a temporal network at the percolation threshold Δt_c. Then, we can study the significance of the nucleus by removing the nodes found in the largest temporal subgraph at $\Delta t = \Delta t_c$ and replicating the percolation studies from the previous Section 4.1. Here, removing the nodes means that we discard all the events of the data set where at least one participant node is marked for removal.

The most descriptive metric to study the effect of removal is the relative size of the largest subgraph in nodes, which we show in Figure 4.4 for the original data and two references with the one-month call data. In the first reference the nodes of the nucleus are removed. In the second, we remove the corresponding number of random nodes. Since the latter reference is stochastic, we average the results over five independent runs. Clearly, the phase transition takes place at larger values of Δt for both references. However, specifically removing the nucleus nodes causes the network to percolate later compared to the random removal of the nodes. Though the difference might not seem enormous, it is good to notice the logarithmic horizontal axis and that the first non-zero value for the targeted removal is when Δt is one day, *i.e.* when the circadian patterns are included.

To conclude, we have shown that the group of nodes forming the nucleus are more influential for the connectivity of the temporal network than a randomly chosen group.

4.3.2 Properties of Nodes in the Nucleus

The next logical step is to study the properties of the nodes in the nucleus and see whether they are different from the rest. Reciprocally, if we find good predictors in terms of local node properties, it enables us to estimate whether a node with given properties would be part of the influential nucleus without complete network information.

We chose to study two temporal and two static node properties which can be calculated locally, *i.e.* using only the events for the specific node or aggregated network information on the node and its first neighbors. The temporal properties we chose

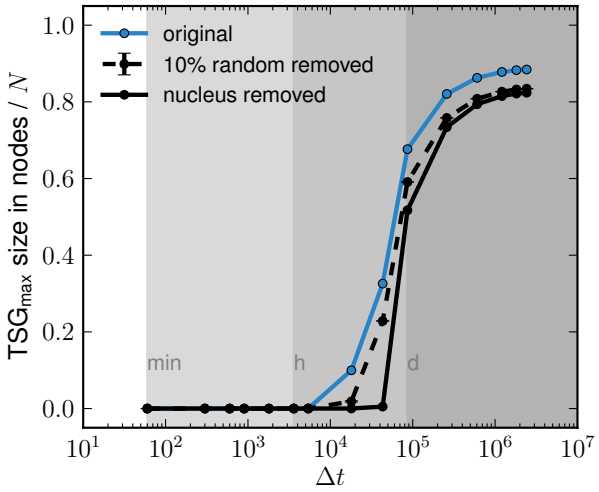


Figure 4.4: The relative size of the $\mathrm{TSG}_{\max}$ for the original network and two references for the one-month call data. In the first reference, the nucleus at $\Delta t = \Delta t_c$ is removed (solid black), and in the second reference a corresponding number (10%) of random nodes are discarded. The latter reference is stochastic, thus we show the average of five independent runs and the standard error of the mean. We see that the targeted removal of the nucleus hinders percolation more than the random removal.

are the total number of events per node E_i and the burstiness of a node B_i. The static metrics are the degree k_i and the clustering coefficient C_i of a node.

In Figure 4.5 we see the relative mean values of these four properties calculated for the nodes in the $\mathrm{TSG}_{\max}$ at a given Δt. Normalization is done with the corresponding network average. At low Δt values, the curves are noisy because the largest subgraphs are small and not yet stable. However, after this initial phase we see some clear trends. The nodes in the largest temporal subgraphs have both larger degree k_i and larger number of events E_i than the nodes who join at larger Δt. The clustering coefficient has more variation and after the stabilization shows a growing trend for the call and SMS networks. In contrast, burstiness has a decreasing trend, especially for the call data. The relative differences are large. For instance in call network the nodes in the nucleus have twice as many neighbors and events than the network average.

Clearly, nodes with large degree, large number of events and high level of burstiness are responsible for the creation of large temporal subgraphs and eventually the nucleus as Δt is increased. However, to get a better view, it is important to also study possible correlations between these properties.

To study the correlations at the node level, we choose two properties and calculate the density of nodes in the network who have specific values for these given properties. In panels (a), (c), and (e) of Figure 4.6 we show the density of nodes given three combinations of the properties for the one-month call data. Immediately we see the strong correlation between the degree and the number of events, as well as

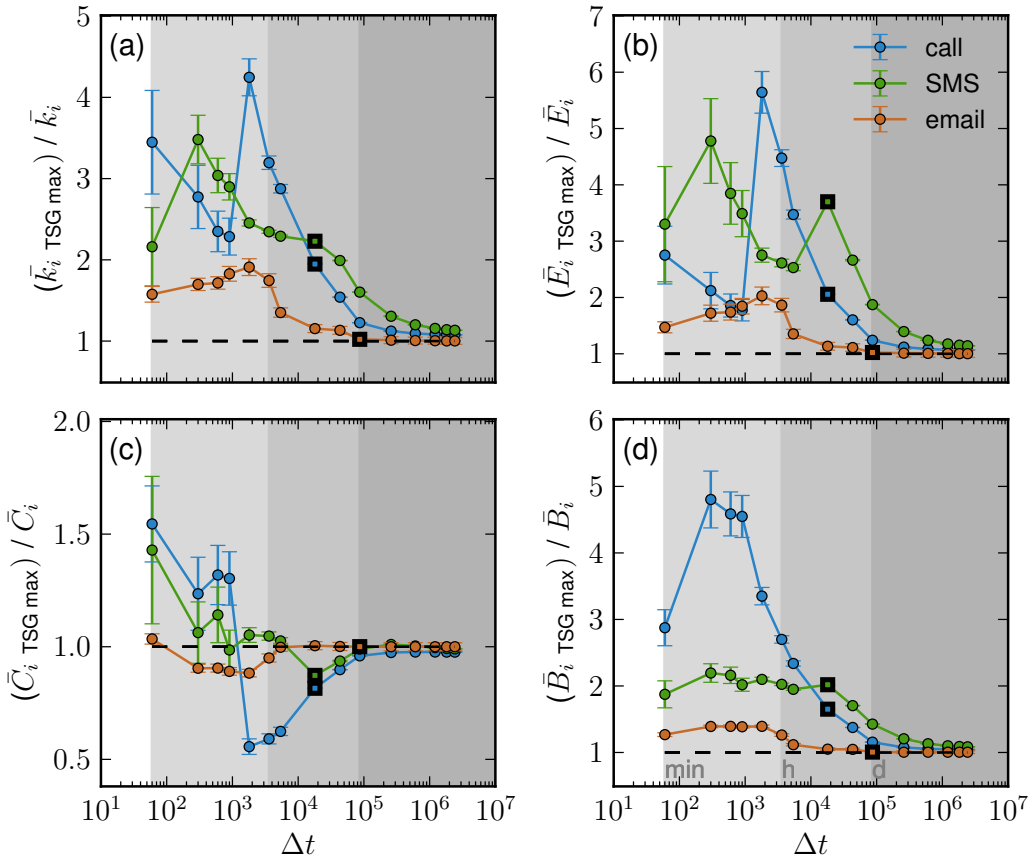


Figure 4.5: The relative mean value of degree k_i (a), number of events E_i (b), clustering coefficient C_i (c) and burstiness B_i (d) of the nodes belonging to the TSG_{max} at a given Δt. Normalization is done with the network average of each property. The $\Delta t = \Delta t_c$ points are emphasized with thick black borders. The error bars show the standard error for each point and the shading emphasizes intervals of one minute, one hour and one day.

degree and burstiness. In contrast, the degree and clustering coefficient are anti-correlated. The other three possible combinations did not show correlations that cannot be inferred from the shown ones.

Some of the observed correlations are trivial. For instance, the anticorrelation between degree and clustering is not surprising, since it is very unlikely that all the neighbors of a high degree node are connected to each other. Also, the correlation between the degree and the number of events is not surprising since people who make many calls are likely to make them to many recipients. What is more interesting is the observed correlation between the degree and the burstiness. As Reference [70] suggests, burstiness is a property of links and not directly a property of nodes except for what is inherited from the links. Then, when the event patterns of the many links of a high degree node are merged to calculate the node burstiness, intuitively one could expect that the burstiness decreases. However, we see the opposite response. One possible explanation for this is the burstiness metric itself,

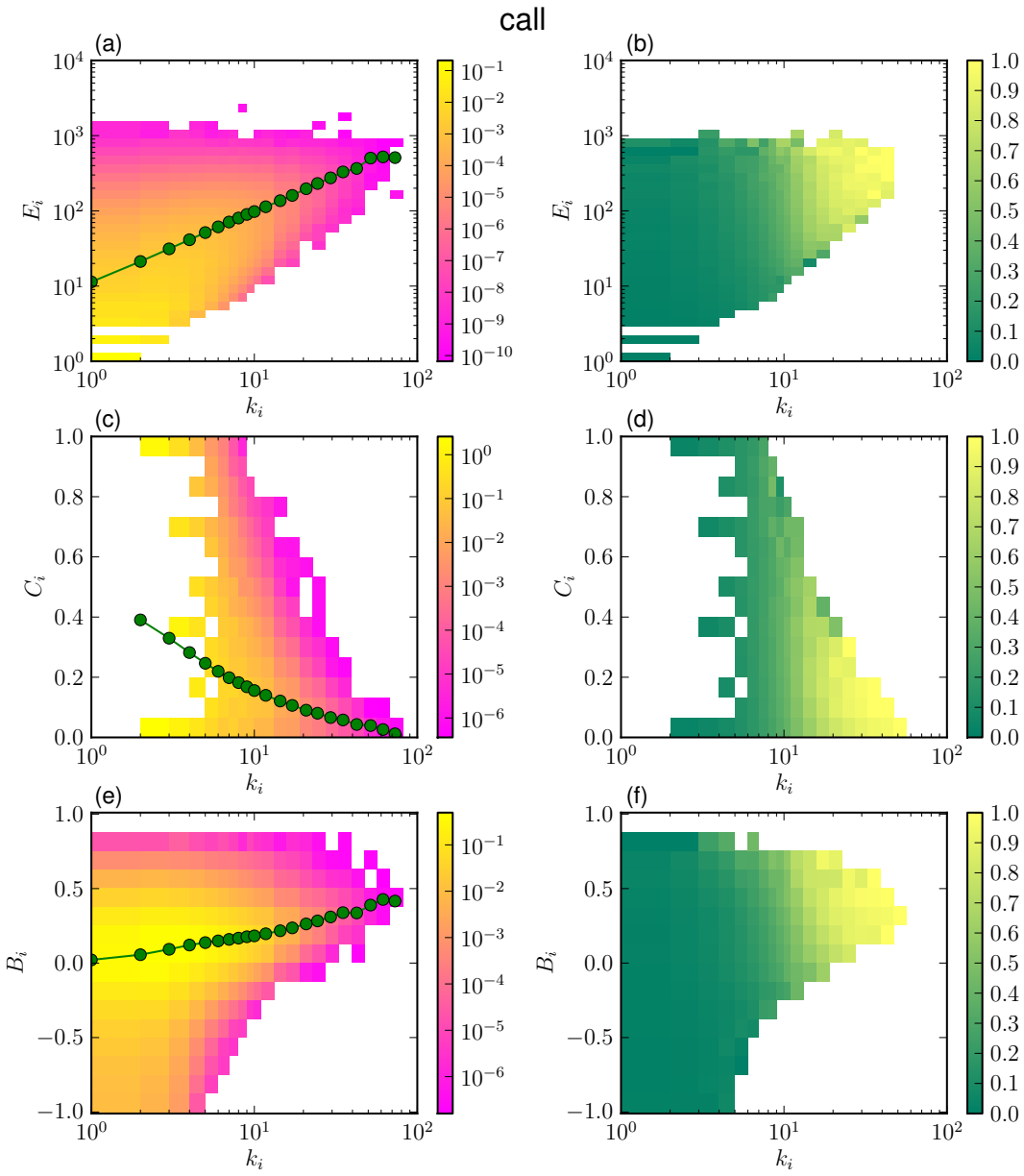


Figure 4.6: Density of nodes with given properties in the one-month call data (panels (a), (c) and (e)) with the average, and the probability that a node with a specific combination of the properties belongs to the $\mathrm{TSG}_{\max}$ at $\Delta t = \Delta t_c$ (panels (b), (d) and (f)). The probability for a bin is shown if at least five nodes fall into it.

since it does not incorporate any component that is affected by the time scale of the event patterns, that is, all distributions with same relationship between the mean and standard deviation result in the same burstiness, even if one pattern spans one month and the other one day. Yet, understanding this phenomenon in depth would require further studies.

In addition to the global correlations between node properties, we are especially interested in how the node properties are correlated when predicting whether a

node belongs to the maximal temporal subgraph. This can be studied by choosing two node properties P_1 and P_2 and studying the probability that a node with a specific combination of these properties belongs to the $\mathrm{TSG}_{\mathrm{max}}$ with given Δt. More formally, we calculate

$$P(\text{in } \mathrm{TSG}_{\mathrm{max}} | P_1 = p_1, P_2 = p_2, \Delta t) = \frac{\#\text{ of nodes in } \mathrm{TSG}_{\mathrm{max}} \text{ with property values } p_1, p_2}{\text{total } \#\text{ of nodes with property values } p_1, p_2}. \quad (4.2)$$

In practice, we use two dimensional binning of the data over the two properties and thus perform the study over small intervals instead of just single values. To study the interesting nucleus, we set $\Delta t = \Delta t_c$ and show the probability in panels (b), (d) and (f) of Figure 4.6. Only bins that have at least 5 nodes with the corresponding property values are shown.

When we compare the densities of the properties and the probability, we see that the nodes who form the nucleus have property values that are rare in the network. Conversely, if we can pick a node with high degree and high number of events, we can be fairly sure that it will belong to the nucleus. When examining the mutual correlations between the properties, we notice that actually the degree of a node is a sufficient predictor for the nodes attendance to the influential nucleus. Results for the SMS and email networks are qualitatively similar and are shown in Appendix A.

4.4 Influential Individual Nodes

In the previous Section we discussed the influence of the nucleus on system-level connectivity and found that degree is a good predictor for a node's attendance to the nucleus. Next, we want to focus closer on the individual nodes, and study their roles within temporal subgraphs. Thus, we'll move on to using the TSGEL method.

From this point on, we concentrate only on the 6-month call data and discard the SMS and email data. The main reason why the call data is more suitable and interesting is that call communication requires mutual activity from both participants of the call. The possible inactivity of message recipients in the other two media hinders the interpretation of possible causal relations between events. Also, a few data-specific features support the rejection of the SMS and email data sets. As already discussed, SMS communication is used mostly in "ping-pong" styled communication which does not promote wider spreading of information. Further, the duration of the email data is short which results in a low number of TSGs per node. Additionally, in both of these media the possibility of sending multimessages and the overall low cost of messaging causes problems when interpreting the results.

First, in order to get started with the TSGEL method, a proper value must be selected for the parameter Δt.

4.4.1 Selecting the Proper Δt Parameter for TSGEL

The upper boundary for the parameter Δt comes directly from the percolation studies. Clearly, if we want the TSGs to represent temporal correlations which happen within a short intervals of each other, the value must be much less than Δt_c. Then what about the lower boundary? What is the value the parameter has to have so that the method has the ability to capture the time scales important for call communication? We can get insight to this by studying the density of preceding events [51, 52].

The density of preceding events is a distribution of intervals between the time of beginning of an event of a node and the end times of all previous events of that same node. Because we are not able to know which events are in reality causally connected, we include all the intervals that fulfill the order condition. When combining the intervals of all the nodes, we see a peak in the distribution at the location of a characteristic interval, if one exists.

In addition, we can categorize the time intervals to preceding events based on the direction of the events and whether two events are between two or three nodes. This gives us altogether eight conditions, denoted with Out $\leftarrow$ In if we calculate the intervals from an outgoing event of a node to the incoming events before it, and Out $\leftarrow_N$ In if we restrict that the outgoing event has to be with a new node, *i.e.* other than the caller of the incoming event. In Figure 4.7 we see the probability density function for the intervals for the call network, categorized in the possible causal intervals in panel (a) and non-causal intervals in panel (b). The maximal interval is restricted to 24 hours. Densities for the RTS reference are also shown.

In panel (a), we see a peak in all distributions at $\sim$ 20 s. This is indicates that it is likely that an individual, after making or receiving a call, makes a new call relatively soon. This phenomenon is in line with the known bursty behavior of humans. The requirement that the later event has a different participant than the earlier one decreases the PDF values for small intervals. This indicates that calls between two participants are more likely than a sequence of calls involving three. The differences in the distributions for very small intervals are most likely due to technical constraints which make calling in certain circumstances slightly faster.

Though the intervals seen in panel (b) do not arise from causal actions of the node for which the specific interval is calculated, they can still represent causal behavior of the other party. This is seen as the disappearing of the characteristic peak when the two incoming calls cannot be from the same node or the incoming call cannot be induced by the outgoing.

The comparison of the distributions with the reference indicates the time scales where either causal or correlated actions affect the inter-event times between the events. Thus, our choice for the parameter Δt must be large enough so that all causality-related peaks are well contained in it. Note that the decrease of the RTS reference near the interval of 24 hours is due to the circadian patterns.

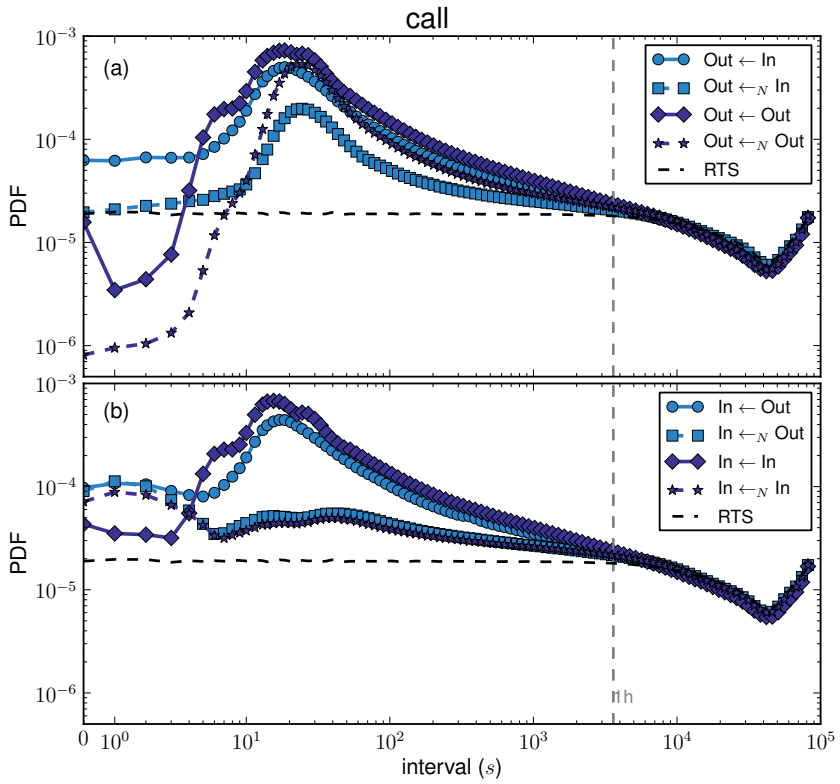


Figure 4.7: Density of preceding events, *i.e.* the probability density function of the time intervals separating events of the nodes. In panel (a) we see time intervals of events with a possible causal relation, whereas in (b) the time intervals are between events that do not have a direct causal relationship. On the left side of the arrow is the direction of the event which happens later in time. The condition $\leftarrow_N$ requires that the latter event is with a new node, *i.e.* other than the other participant in the earlier event. The horizontal dashed line is the RTS reference, and the vertical dashed line emphasizes the 1-hour interval.

Based on these motivations, we choose the value $\Delta t = 3600$ s for the TSGEL studies. Then, 33.7% of the TSGs have at least 2 events and 16.5% at least 3 nodes (the absolute number is $\sim 52 \times 10^6$). The size of the TSG_{max} is 2591 nodes (6780 events).

4.4.2 Role of the Nodes within a Subgraph

After running the TSGEL method with $\Delta t = 3600$ s, each root and tip of the events that are connected to a given node has a label which illustrates how the node has acted with respect to other events. By summing the labels of a given node, we can study what role the node on average has in temporal subgraphs. We denote the sum of the source labels of node i by $\sigma_{i,\text{SRC}}$ and the sum of the sink labels by $\sigma_{i,\text{SNK}}$. We discard subgraphs with less than three unique nodes from the analysis since we are interested in node roles within local neighborhoods, and not in reciprocal communication with only one other node.

The sum of the event labels per node are trivially correlated with the number of the node's events. More specifically, the source and transmitter-out scores are equal to the number of outgoing events, and the sink and transmitter-in scores sum up to the number of incoming events. Since the sizes of the TSGs are relatively small with the chosen Δt, the sum of sink and source scores should be closely related to the number of the outgoing and incoming events, respectively. This is exactly what we observe in panels (a) and (b) of Figure 4.8, which shows the density plot of the number of outgoing events $E_{i,\mathrm{OUT}}$ versus the source score $\sigma_{i,\mathrm{SRC}}$, and the corresponding numbers for the incoming events and sink scores. The average of the distribution is close to the linear upper bound, indicating that most of the outgoing events begin only after a time span Δt from incoming event, and reciprocally that an incoming event is not likely to spark up new events within the specified interval. This observed symmetry comes directly from the TSGEL method: for instance, if an outgoing event would be a transmitter event and thus make the source label count smaller, it would have to have at least one matching incoming event which would also be a transmitter event instead of a sink event.

However, the small differences between the label sums from the one-to-one match between the event counts are interesting. For instance, are some node's events more likely to be transmitters than the events of some other node? If so, this reveals that according to the event correlations, the node plays the role of a transmitter in its neighborhood. We measure this with the relative source score $\sigma_{i,\mathrm{SRC}}/E_{i,\mathrm{OUT}}$ for each node and study how it varies as a function of the four properties chosen previously, namely the number of events, degree, clustering coefficient and burstiness of a node. The conditional probability of the relative source score as a function of a given property is seen in panels (c)-(f) of Figure 4.8. Note that because of the symmetry, the results would agree if we would instead study the relative sink score.

We observe that the likelihood of an event having a source label decreases as a function of the number of outgoing events, degree and burstiness. In other words, nodes with high values in these metrics are likely to have an incoming call taking place close enough before they make a call. The decrease in the likelihood is significant, for example, low-degree nodes have source events with $\sim 80\%$ chance whereas very high degree nodes have transmitter labels more than half of the time. The fourth property, *i.e.* the clustering coefficient, does not explain the relative source score.

The results on the roles of nodes within temporal subgraphs are in perfect agreement with the results seen with the node properties of the nucleus in Section 4.3. Actually, since the event labeling method with $\Delta t < \Delta t_c$ reveals the behavior of a node within its topologically local neighborhood and temporally nearby events, the observed behavior of the high number of events, high degree and high burstiness nodes indicates that eventually these nodes join together and form the nucleus. Thus, the nodes who act as transmitters and do not stop the growth of the temporal subgraphs are influential with respect to temporal connectivity in the system. Clearly, the underlying correlations between the properties seen in Figure 4.6 do still play a role here.

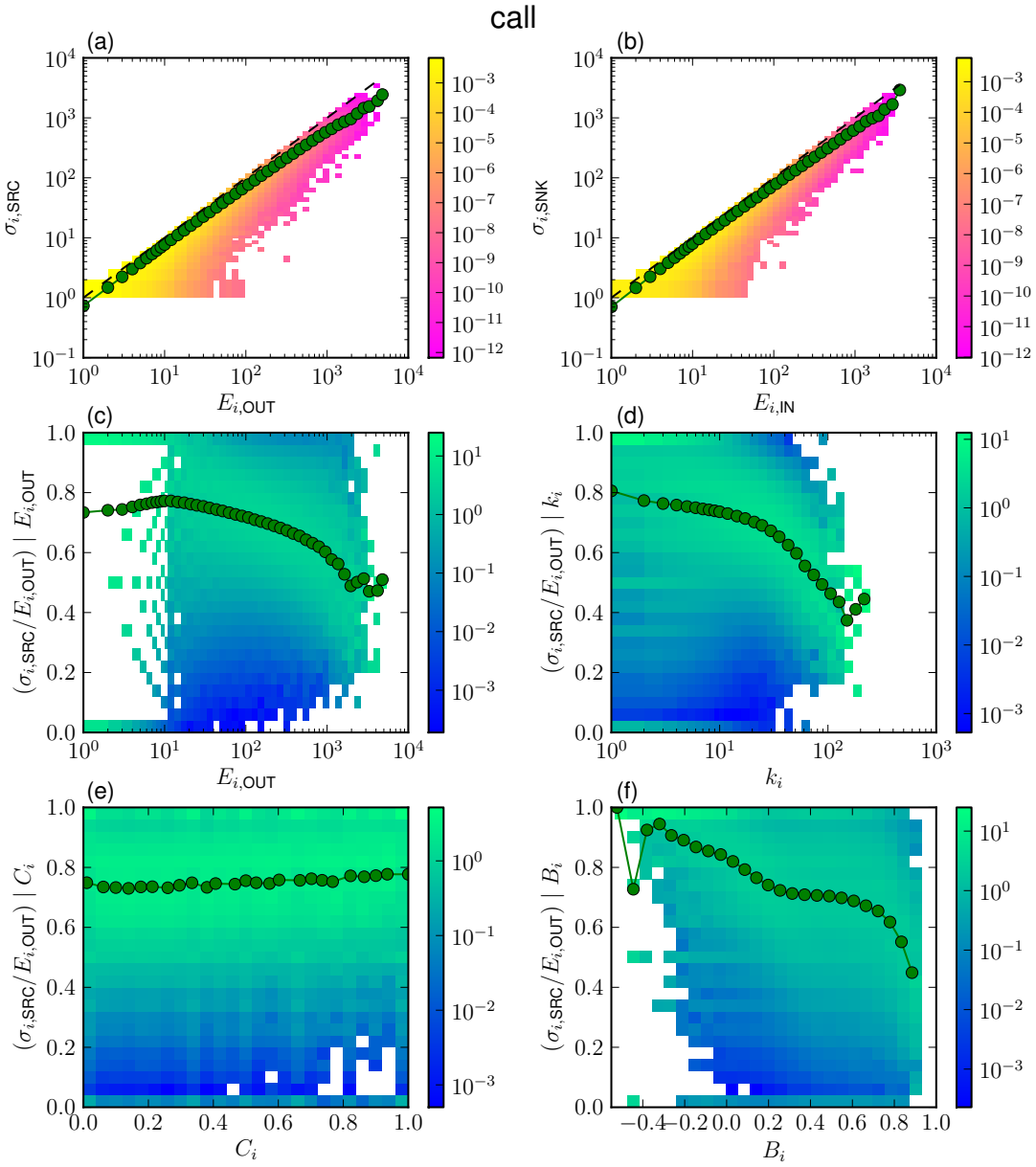


Figure 4.8: Roles of nodes in temporal subgraphs for the 6-month call data with $\Delta t = 3600$ s. Only TSGs with at least three nodes, and nodes with at least 10 TSG participations are taken into account. In panels (a) and (b) we show the density plot for the number of events of a node versus the sum of corresponding labels. Panels (c)-(f) show the conditional probability of the relative source number against node properties (the number of out-events, degree, clustering coefficient and burstiness of a node). The green circles show bin averages in all panels.

4.4.3 Size of the TSG a Node Generates

Finally, we will study how large temporal subgraphs nodes sharing a property on average create. We study this by first choosing a node property, then taking a subset of the temporal subgraphs where a node with given property acts as a source at least once, and finally calculating the conditional distribution of the size of the TSG as a function of the given property. As above, we require that the TSGs have at least three nodes.

In panel (a) of Figure 4.9 we show the conditional distribution of the TSG size, measured in nodes, as a function of the number of events of a node. Clearly, as the number of events grow, the larger temporal subgraphs a node creates. Or more specifically, *joins*, since the node is not necessarily the one that initiates the subgraph. Basically, what is seen is related to the result in Section 4.3.2 that nodes with high numbers of events participate in large subgraphs. However, as temporal subgraphs can have multiple sources, looking at their size alone can be misleading

To get a better view of the size of the temporal subgraph that results from the actions of a single source, we scale the size of the TSG with the number of sources it has. If a single node has multiple source events in a single subgraph, it is counted as one in the scaling. The conditional distribution of scaled size as a function of number of events is seen in panel (b) of the Figure 4.9. As expected, the size plummets when comparing to the unscaled metric. However, we observe a small but clear increasing trend in the scaled size of the TSG. Though the absolute increase is not large, crossing the TSG size of two can have significant consequences considering the percolation.

Consider a case where the source contacts one node, resulting in a TSG of size two. Then, in order that the size of the TSG would be larger than two, the initial node must contact a third node, or the first contacted node must contact a new node. Clearly, this results in a better starting point for system-wide percolation of the TSG (see the reproductive number in *e.g.* [8]). In the other end, nodes with average scaled TSG sizes less than two are connected to nodes behaving like sinks, *i.e.* they do not contact new nodes even when they get calls from multiple unique sources.

As expected, the outcomes as a function of degree and burstiness agree with the ones obtained with the number of events and thus they are not reported here. Also, as above, clustering does not correlate with the TSG size. We also replicated the results with $\Delta t = 300$ s. Then, naturally, the raw size metric is much smaller even with large property values, but the interesting observation of crossing the TSG size of two with the scaled metric is still present.

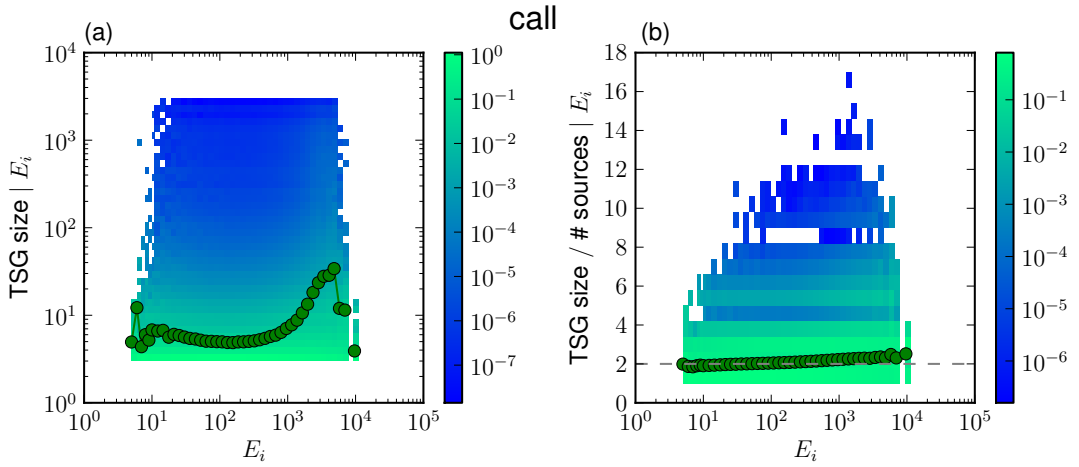


Figure 4.9: The conditional distribution of the size of the TSG a source node participates in, as a function of the number of events of the source node. Only TSGs with at least three nodes and nodes with at least 5 TSG participations as a source are taken into account. Panel (a) shows the size of the TSG in unique nodes, whereas in panel (b) the size is scaled with the number of unique source nodes.

4.5 Discussion on the Influential Nodes

In the last two Sections we presented results on the second objective of this Thesis, namely how can we recognize influential nodes in temporal communication networks. We approached the question at two levels. First, we observed that the nucleus, the set of nodes around which a giant temporal subgraph emerges, is influential when it comes to connectivity of the system, and then studied properties of the nodes forming the nucleus. We found that nodes with high degree, high burstiness and large number of events are likely to be in the nucleus, and are thus influential for connectivity. It was also seen that the degree, burstiness and number of events are highly correlated between each other and the fourth property, clustering coefficient, is anticorrelated with all of them. Then we used the TSGEL method to study how nodes' actions influence its local neighborhood. We found out, again, that nodes with high degree, high burstiness and large number of events are rarely the ones where the growth of a temporal subgraph finishes (*i.e.* sinks); rather, subgraphs where such nodes act as sources are on average large. When the result is viewed in the context of information diffusion, nodes with these properties are important in their local neighborhood for transmitting information onward.

Results at both the group level (nucleus) and the individual level support each other perfectly. To summarize, nodes with high degree and large number of events reach more nodes when they act as sources. Additionally, nodes with high values of these properties are unlikely to act as sinks that stop TSG growth. Thus, they are likely to participate in large and long-living subgraphs, and eventually form the nucleus.

It is also important to discuss some of the features of the TSGEL method. Even though the labeling is based on assumed causality of the events, the temporal sub-

graphs are not necessarily causal. This was seen for instance when studying the size of the TSG a source node creates: larger TSGs have nearly always multiple sources who are not causally connected. Thus, the method is essentially able to only reveal correlations within a node's neighborhood: a node gets many transmitter labels if it is in a neighborhood with many temporally adjacent events. This becomes problematic with nodes that have a high number of events: no matter how their events are distributed, they can't be silent for long periods of time, and thus they are rarely sources or sinks. Therefore, one could consider using different values of the parameter Δt for different nodes, reflecting their level of activity. On the other hand, isn't the fact that high-activity nodes are not likely to remain inactive for any longer periods of time exactly what makes them influential?

One possible approach that would enable us to make claims about the causal influence of a node would be to study causal paths or reachability within a temporal subgraph. However, we would still need to make assumptions that events respecting the direction and time are causal if they happen close enough to each other. On the basis of this work and the problems encountered, the authors suggest that if one wants to make conclusions on the causal influence of a node, one would need to use data where such information is present (see *e.g.* Ref. [66]).

Our analysis of the influential nodes was mostly descriptive. However, to be able to make claims on the unexpectedness of the results, we would need a reference to compare with. For example, we could think of breaking the correlations between degree, number of events and burstiness to see which of these has the largest effect. However, it must be noted that then we would also break correlations that are clearly characteristic of the individuals that form the network. We would end up with a reference, but one can question the meaningfulness of a reference where characteristic features of human behavior have been erased.

5 Summary and Conclusions

Temporal networks research is currently an active field because of theoretical advances and especially the availability of suitable data sets. In this Thesis we have studied three empirical communication networks, with two main objectives. The first was to study temporal percolation in communication networks, and the second to study whether we are able to identify which nodes in temporal networks are the most influential regarding temporal connectivity and flow of information.

We observed a percolation transition in all the networks, in terms of temporal subgraphs suddenly spanning the entire network over its entire lifetime. We were also able to approximately determine the corresponding percolation thresholds, that is, critical values of the temporal adjacency parameter used in constructing temporal subgraphs. Because of the additional dimension of time, there are issues to be considered beyond static percolation theory, and special attention was paid to these. We found out that the concept of the lifetime of temporal subgraphs is important when defining the percolation threshold. With the lifetime, we were able to determine the nucleus of the network – an influential group of nodes that ultimately form the giant component. As the giant component covers most of the nodes, the nucleus takes care of covering the time dimension. The percolation threshold has an important consequence when considering dynamical processes on temporal networks: any process with a characteristic time scale of node deactivation less than the percolation threshold cannot reach a significant proportion of the nodes or cover a significant fraction of the data time span of the network.

The second objective was split into two parts. First, we studied the properties of the nodes that form the influential nucleus. We concentrated on node properties that can be calculated from local information on the nodes only. We found out that the degree, number of events and burstiness are highly correlated, and that the nodes with a high value in these properties are likely to belong to the nucleus. Second, we introduced a method for labeling the events within a temporal subgraph and found out that, again, nodes with high degree, large number of events and high burstiness are significant on two aspects: first, they rarely cease the growth of a TSG and second, when acting as sources, they initiate slightly larger TSGs.

To conclude, the lifetime of a temporal subgraph is an important concept for temporal percolation transition. Also, even though the problem of significant nodes can be approached in multiple different ways – many more than discussed in this Thesis – it is safe to assume that nodes with large degree and large number of events are usually influential in a temporal network, and responsible for temporal percolation transitions.

References

[1] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, pp. 651–4, Oct. 2000.

[2] B. A. Huberman, *The laws of the Web: Patterns in the ecology of information*. MIT Press, 2003.

[3] W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[4] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[5] S. Wasserman and K. Faust, *Social Network Analysis: Methods and applications*. Cambridge University Press, 1994.

[6] R. Bond, C. Fariss, J. Jones, and A. Kramer, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, pp. 295–8, Sept. 2012.

[7] P. Holme and J. Saramäki, "Temporal Networks," *Physics Reports*, vol. 519, pp. 97–125, Oct. 2012.

[8] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, USA, 2010.

[9] M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review*, vol. 45, pp. 167–256, Sept. 2003.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks.," *Nature*, vol. 393, no. 6684, pp. 440–2, 1998.

[11] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters*, vol. 89, p. 208701, Oct. 2002.

[12] A.-L. Barabási, "Emergence of Scaling in Random Networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, pp. 661–703, Nov. 2009.

[14] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proceeding of the 17th international conference on World Wide Web - WWW '08*, (New York, NY, USA), p. 915, ACM, Apr. 2008.

[15] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks.," *Science*, vol. 298, pp. 824–7, Oct. 2002.

[16] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75–174, Feb. 2010.

[17] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 7821–6, June 2002.

[18] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 1118–23, Jan. 2008.

[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, Oct. 2008.

[20] D. Hric, R. K. Darst, and S. Fortunato, "Community detection in networks: structural clusters versus ground truth," *arXiv preprint arXiv:1406.0146*, p. 21, 2014.

[21] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[22] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[23] M. E. J. Newman, "Spread of epidemic disease on networks," *Physical Review E*, vol. 66, p. 016128, July 2002.

[24] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, May 2005.

[25] M. Karsai, M. Kivelä, R. K. Pan, and K. Kaski, "Small but slow world: How network topology and burstiness slow down spreading," *Physical Review E*, vol. 83, p. 025102, 2011.

[26] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, "Burstiness and spreading on temporal networks," *The European Physical Journal B*, vol. 86, p. 320, May 2013.

[27] K.-I. Goh and A.-L. Barabási, "Burstiness and memory in complex systems," *Europhysics Letters (EPL)*, vol. 81, p. 48002, 2008.

[28] D. Kempe, J. Kleinberg, and A. Kumar, "Connectivity and Inference Problems for Temporal Networks," in *Journal of Computer and System Sciences*, vol. 64, pp. 504–513, 2000.

[29] P. Holme, "Network reachability of real-world contact sequences," *Physical Review E*, vol. 71, p. 046119, Apr. 2005.

[30] R. K. Pan and J. Saramäki, "Path lengths, correlations, and centrality in temporal networks," *Physical Review E*, vol. 84, p. 016105, July 2011.

[31] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, "Temporal motifs in time-dependent networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, p. P11005, Nov. 2011.

[32] L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki, "Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 45, pp. 18070–18075, 2013.

[33] R. Pfitzner, I. Scholtes, A. Garas, C. J. Tessone, and F. Schweitzer, "Betweenness Preference: Quantifying Correlations in the Topological Dynamics of Temporal Networks," *Physical Review Letters*, vol. 110, p. 198701, May 2013.

[34] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, "Activity driven modeling of time varying networks," *Scientific Reports*, vol. 2, p. 469, Jan. 2012.

[35] M. Starnini and R. Pastor-Satorras, "Temporal percolation in activity-driven networks," *Physical Review E*, vol. 89, p. 032807, 2014.

[36] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network Analysis in the Social Sciences," *Science*, vol. 323, pp. 892–5, 2009.

[37] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[38] M. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[39] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 7332–6, May 2007.

[40] R. B. Cialdini and N. J. Goldstein, "Social influence: compliance and conformity.," *Annual Review of Psychology*, vol. 55, pp. 591–621, Jan. 2004.

[41] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, vol. 311, pp. 854–6, 2006.

[42] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, vol. 329, pp. 1194–7, Sept. 2010.

[43] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The Role of Social Networks in Information Diffusion," in *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528, ACM, 2012.

[44] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," *Science*, vol. 337, no. 2012, pp. 337–341, 2012.

[45] J. Moody, "The Importance of Relationship Timing for Diffusion," *Social Forces*, vol. 81, pp. 25–56, Sept. 2002.

[46] P. Holme and F. Liljeros, "Birth and death of links control disease spreading in empirical contact networks.," *Scientific Reports*, vol. 4, p. 4999, 2014.

[47] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks.," *Journal of theoretical biology*, vol. 271, pp. 166–180, Dec. 2010.

[48] J. A. Danowski and P. Edison-Swift, "Crisis Effects on Intraorganizational Computer-Based Communication," *Communication Research*, vol. 12, pp. 251–270, Apr. 1985.

[49] G. Kossinets, J. Kleinberg, and D. J. Watts, "The structure of information pathways in a social communication network," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, (New York, NY, USA), p. 435, ACM Press, Aug. 2008.

[50] J.-P. Eckmann, E. Moses, and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 14333–7, Oct. 2004.

[51] V.-P. Backlund, J. Saramäki, and R. K. Pan, "Effects of temporal correlations on cascades: Threshold models on temporal networks," *Physical Review E*, vol. 89, p. 062815, June 2014.

[52] L. Kovanen, "Structure and dynamics of a large-scale complex social network," Master's thesis, Aalto University, Finland, 2009.

[53] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?," in *Proceedings of the 19th international conference on World Wide Web*, pp. 591–600, ACM, 2010.

[54] J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "Locating privileged spreaders on an online social network," *Physical Review E*, vol. 85, p. 066123, June 2012.

[55] S. A. Myers and J. Leskovec, "The bursty dynamics of the Twitter information network," in *Proceedings of the 23rd international conference on World Wide Web*, pp. 913–924, ACM, 2014.

[56] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer Networks," *Journal of Complex Networks*, vol. 2, pp. 203–271, 2014.

[57] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, "Measuring Large-Scale Social Networks with High Resolution.," *PLoS ONE*, vol. 9, p. e95978, Jan. 2014.

[58] M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai, "Multiscale analysis of spreading in a large communication network," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, p. P03005, Mar. 2012.

[59] J. Saramäki, E. A. Leicht, E. López, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar, "Persistence of social signatures in human communication.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 942–947, Jan. 2014.

[60] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[61] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, p. 35, 1977.

[62] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "k-Core Organization of Complex Networks," *Physical Review Letters*, vol. 96, p. 040601, Feb. 2006.

[63] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, pp. 888–893, Aug. 2010.

[64] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras, "Immunization strategies for epidemic processes in time-varying contact networks," *Journal of Theoretical Biology*, vol. 337, pp. 89–100, 2013.

[65] S. Liu, N. Perra, M. Karsai, and A. Vespignani, "Controlling Contagion Processes in Activity Driven Networks," *Physical Review Letters*, vol. 112, p. 118702, Mar. 2014.

[66] S. Pei, L. Muchnik, J. S. Andrade, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media.," *Scientific Reports*, vol. 4, no. c, p. 5547, 2014.

[67] J. Ruths and D. Ruths, "Control Profiles of Complex Networks," *Science*, vol. 343, pp. 1373–6, Mar. 2014.

[68] B. Ribeiro, N. Perra, and A. Baronchelli, "Quantifying the effect of temporal resolution on time-varying networks," *Scientific Reports*, vol. 3, p. 3006, 2013.

[69] G. Krings, M. Karsai, S. Bernhardsson, V. D. Blondel, and J. Saramäki, "Effects of time window size and placement on the structure of an aggregated communication network," *EPJ Data Science*, vol. 1, p. 4, May 2012.

[70] M. Karsai, K. Kaski, and J. Kertész, "Correlated Dynamics in Egocentric Communication Networks," *PLoS ONE*, vol. 7, no. 7, p. e40612, 2012.

A Node Properties in the Nucleus of the SMS and Email Data

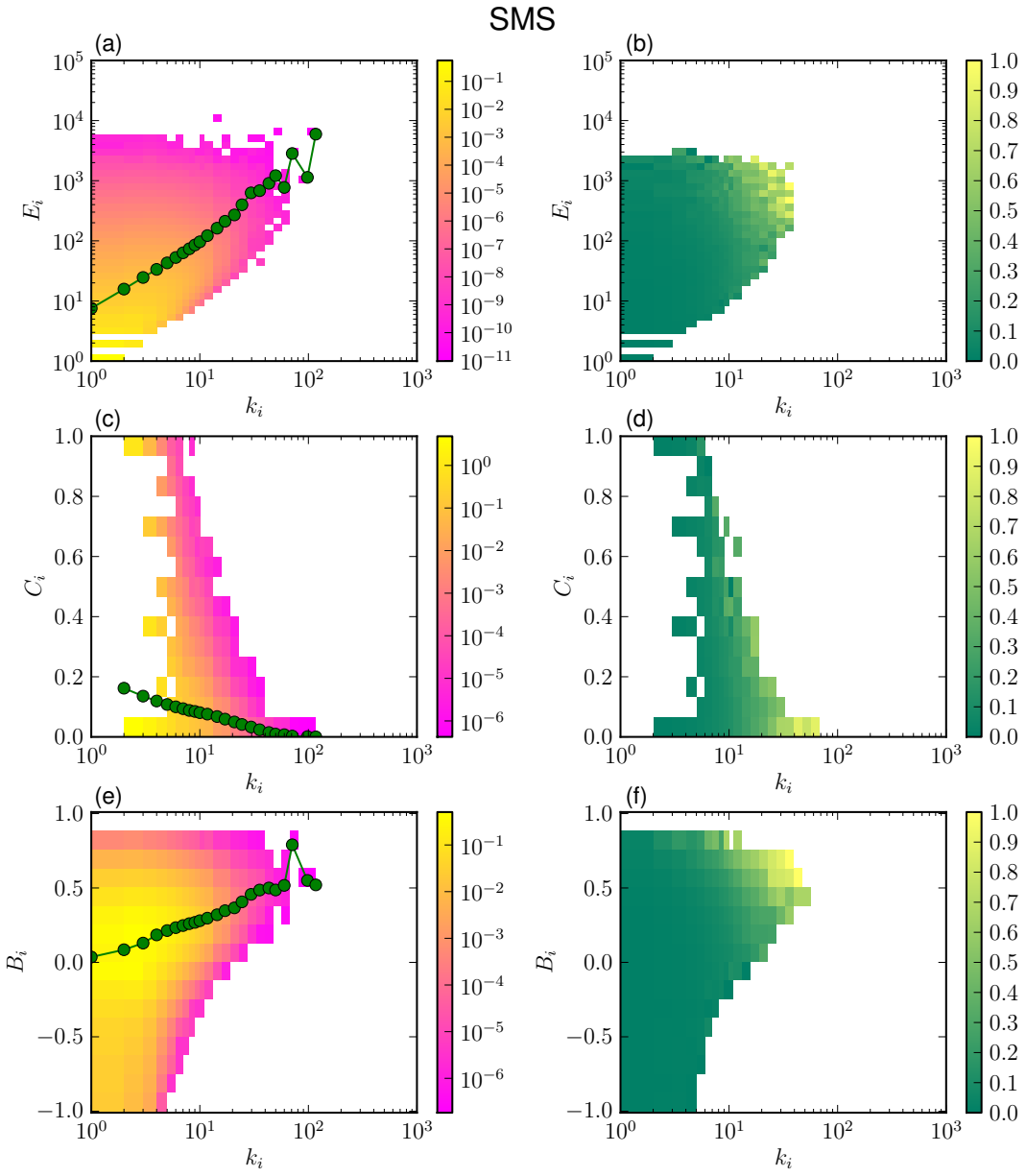


Figure A1: Density of nodes with given properties in the SMS data (panels (a), (c) and (e)) with the average, and the probability that a node with a specific combination of the properties belongs to the $\mathrm{TSG}_{\max}$ at $\Delta t = \Delta t_c$ (panels (b), (d) and (f)). The probability of a bin is shown if at least five nodes fall into it.

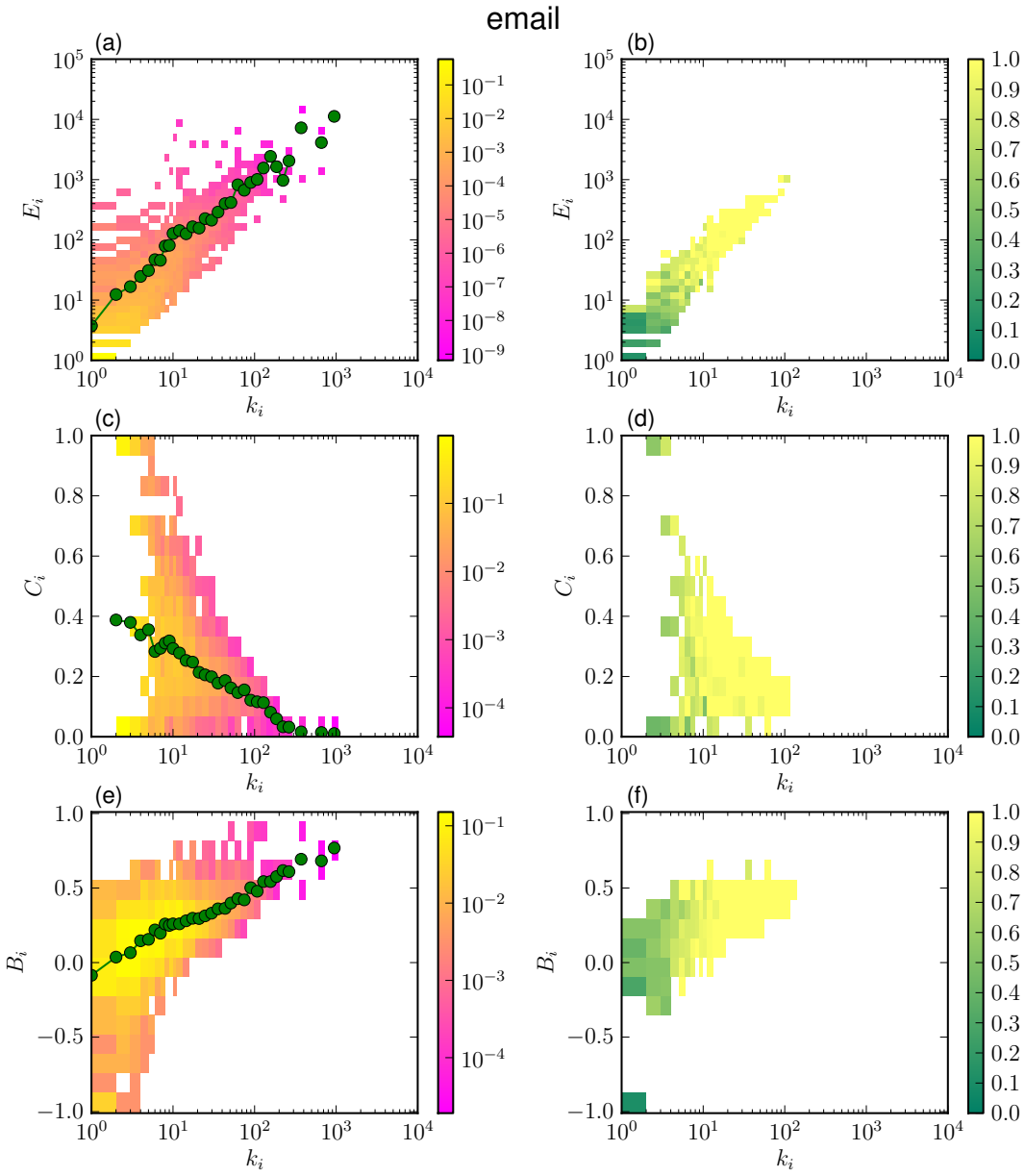


Figure A2: Density of nodes with given properties in the email data (panels (a), (c) and (e)) with the average, and the probability that a node with a specific combination of the properties belongs to the $\mathrm{TSG}_{\max}$ at $\Delta t = 5400$ s (panels (b), (d) and (f)). The probability of a bin is shown if at least five nodes fall into it. Note that for the email network we chose a much smaller value for Δt than the percolation threshold, since at Δt_c nearly all the nodes are in the $\mathrm{TSG}_{\max}$ and the probability would just be 1.